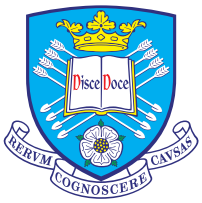
# Novel Matrix Completion Methods for Missing Data Recovery in Urban Systems




The University Of Sheffield.

**Cristian Genes**

Department of Automatic Control and Systems Engineering
University of Sheffield

This dissertation is submitted for the degree of
*Doctor of Philosophy*

March 2019

# Acknowledgements

I would express my gratitude to my advisor Dr. Iñaki Esnaola for his guidance and continuous support of my Ph.D study. His patience, motivation and immense knowledge helped me in all the time of research and writing of this thesis.

My sincere thanks also goes to my advisor Prof. Daniel Coca for his insightful comments and encouragement and also for giving me the opportunity to join the project.

I would like to thank Dr. Samir M. Perlaza and Prof. Nando Ochoa for the stimulating discussions and for sharing their expertise. Special mention to my thesis examinators who provided constructive feedback for my thesis and research.

This dissertation would not have been possible without funding from the University of Sheffield Future Cities Scholarship.

I would like to thank my family for supporting me throughout my Ph.D studies and in my life in general.

# Abstract

Urban systems are composed of multiple interdependent subsystems, for example, the traffic monitoring system, the air quality quality monitoring system or the electricity distribution system. Each subsystem generates a different dataset that is used for analyzing, monitoring, and controlling that particular component and the overall system. The quantity and quality of the data is essential for the successful applicability of the monitoring and control strategies. In practical settings, the sensing and communication infrastructures are prone to introduce telemetry errors that generate incomplete datasets. In this context, recovering missing data is paramount for ensuring the resilience of the urban system.

First, the missing data from each subsystem is recovered using only the available observations from that subsystem. The fundamental limits of the missing data recovery problem are characterized by defining the optimal performance theoretically attainable by any estimator. The performance of a standard matrix-completion based algorithm and the linear minimum mean squared error estimator are benchmarked using real data from a low voltage distribution system. The comparison with the fundamental limit highlights the drawbacks of both methods in a practical setting. A new recovery method that combines the optimality of the Bayesian estimation with the matrix completion-based recovery is proposed. Numerical simulations show that the novel approach provides robust performance in realistic scenarios.

Secondly, the correlation that results from the interdependence between the subsystems of an urban system is exploited in a joint recovery setting. To that end, the available observations from two interconnected components of an urban system are aggregated into a data matrix that is used for the recovery process. The fundamental limits of the joint recovery of two datasets are theoretically derived. In addition, when the locations of the missing entries from each dataset are uniformly distributed, theoretical bounds for the probability of recovery are established and used to minimize the acquisition cost. The numerical analysis shows that the proposed algorithm outperforms the standard matrix completion-based recovery in various noise and sampling regimes within the joint recovery setting.

# Contents

# List of Figures

# List of Tables

# Notations

In this thesis, matrices are denoted with uppercase bold letters, i.e., $\mathbf{M}$, and vectors are denoted with lowercase bold letters, i.e., $\mathbf{m}$. The superscript $T$ stands for transposition. For matrix $\mathbf{M}$, $\text{tr}(\mathbf{M})$ denotes its trace and $\text{rank}(\mathbf{M})$ denotes its rank. The Frobenius norm is denoted by $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_*$ denotes the nuclear norm of the matrix $\mathbf{M}$. The Euclidean inner product between the matrices $\mathbf{A}$ and $\mathbf{M}$ is denoted by $\langle \mathbf{A}, \mathbf{M} \rangle$. The logarithm function is assumed to be in base 2 unless specified otherwise.

**Acronyms**

AWGN Additive White Gaussian Noise

BSVT Bayesian Singular Value Thresholding

ENWL Electricity North West Limited

HPC High Performance Computing

LMMSE Linear Minimum Mean Squared Error

LV Low Voltage

MIDAS Motorway Incident Detection and Automatic Signalling

MMSE Minimum Mean Squared Error

MSE Mean Squared Error

NMSE Normalized Mean Squared Error

OPTA Optimal Performance Theoretically Attainable

SMR Signal-to-Mismatch Ratio

SNR Signal-to-Noise Ratio

SURE Stein's Unbiased Risk Estimate

SVT Singular Value Thresholding

# Chapter 1

# Introduction

## 1.1 Background and motivation

The United Nation Population Division reported in 2009 that the total number of people living in urban and metropolitan areas exceeded that of those living in rural areas [1] and [2]. In addition, 66% of the world's population will live in urban areas [3] by 2050, which gives rise to challenges regarding air quality, congestion, waste management and human health [4]. In an attempt to tackle these challenges, the interest in the smart city concept has seen a significant growth [5]. This is manifested through a high number of smart city initiatives on city implementation projects or jointly funded public research projects. In [6] it is reported that in 2012 there were 143 ongoing smart city projects of which 47 in Europe and 30 in the USA.

The concept of a smart city is defined in [7] as an instrumented, interconnected and intelligent city. Instrumentation enables the collection of real-world data through the use of sensors, meters, personal devices, cameras, the web and other data-acquisition systems. Interconnection refers to the integration of all types of data into a computing platform that allows the communication between different city services. Finally, intelligence means the use of data analysis tools, modeling and optimization techniques that provide additional insight for the decision makes [7]. In view of this, a smart city relies, among others, on a collection of computing technologies applied to infrastructure components [8]. To facilitate that, datasets are produced automatically and routinely from sensors and other data-acquisition systems [9]. The use of data analysis tools, modeling and optimization techniques relies on the quantity and the quality of the data collected by the sensing infrastructures of the city. Ensuring timely and accurate measurements is key for the managing and controlling the urban system.

For example, the lack of data quality in power systems contributed towards several large-scale blackouts [10] such as the 2003 U.S.-Canadian blackout [11], the 2003 Italy blackout [12], the 2011 Arizona-southern California blackout [13], and the

2012 Indian blackout [14], which resulted in significant economic losses and social impacts. Therefore, it is crucial to ensure the quality of the datasets produced by the sensing infrastructures covering urban areas.

In [15], the quality of the data is influenced by three factors:

- precision of the data collection devices,
- quality of the data communication infrastructure,
- granularity of the measurements in both spatial and temporal dimensions.

In addition to the limitations imposed by the hardware technology used in the acquisition process, the quality of the data is affected by external factors like false data injection attacks or missing data. In false data injection attacks an adversary hacks the readings of multiple sensors to mislead the decision making process [16]. For the electricity grid, the problem of estimating the state of the system under data injection attacks is studied in [17] and [18]. On the other hand, missing data is caused by lost measurements in the acquisition process. Sensor failures, unreliable communication or data storage issues are some of the causes for incomplete sets of observations [19]. As a consequence, the state of the system is not perfectly known and data analysis tools, modeling and optimization techniques are difficult to implement. For instance, accurate measurements are necessary to implement a centralized control scheme for voltage regulation in distribution systems [20]. In view of this, it is vital to develop estimation procedures for the missing data using the available observations. Note that in a missing data recovery framework, the quality of the available observations is affected by the three factors identified in [15]. Therefore, the data recovery techniques need to be robust to noise in the available observations, sampling patterns and mismatch in any prior knowledge incorporated.

Characterizing the fundamental limits of the missing data recovery problem is essential for understanding the trade offs between different recovery techniques. Moreover, when the performance of the state-of-the-art recovery methods is largely suboptimal compared to the theoretical limit or the assumptions are not feasible in a realistic scenario, new recovery strategies need to be developed to overcome those limitations. In the smart grid context, the problem of estimating the state of the system from partial observations is addressed in [21], [19] and [22].

Urban systems are typically modelled as a set of interacting subsystems [23]. One of the consequences of the interaction between different subsystems is that there is correlation between the datasets generated by each component. In view of this, the missing data recovery performance for one subsystem can be enhanced by using available observations collected by other interacting components of the urban system. A joint estimation framework enables the recovery when the number of available observations from one dataset is not sufficient by compensating with

an increased number of observations from the other datasets. The joint recovery paradigm supports an additional type of correlation to be exploited in the recovery process. Specifically, the interaction between different components of the urban system gives rise to correlation between the datasets collected by each sensing infrastructure. To decide when is beneficial to do joint missing data recovery for multiple datasets, it is vital to characterize the fundamental limits. On the practical side, the recovery algorithms need to be robust to different sampling regimes in the datasets combined.

To sum up, the main challenges faced by the missing data recovery strategies are:

- uncertainty in the quality of the available observations,
- randomness in the number and the location of the missing data,
- mismatch in any prior knowledge incorporated.

In view of this, developing robust recovery methods is paramount for ensuring the quality and quantity of the data in urban systems.

## 1.2 Overview of the thesis

The rest of the thesis addresses the problem of recovering missing data from datasets generated in urban systems.

- Chapter 2 presents a review of the state-of-the-art techniques for recovering missing data particularly focusing on a new research field called matrix completion.
- Chapter 3 defines a framework for comparing the recovery performance of different estimation methods using real data collected from a low voltage distribution system. Moreover, the fundamental limits of the recovery process are characterized when the optimal performance theoretically attainable is defined within an information theoretic framework. A standard matrix-completion-based algorithm is compared with the linear minimum mean squared error estimator in a practical setting. The performance limitations of both recovery methods relative to the fundamental limit are identified. The findings of this chapter are published in [19].
- Chapter 4 introduces a novel algorithm for recovering missing data that addresses the key performance limitations of the methods compared in Chapter 3. The numerical analysis demonstrates the improved accuracy and the robustness of the proposed approach with respect to noise in the available observations, non-uniform sampling patterns and mismatch in the prior knowledge. The

results in this chapter are published in the *IEEE Trans. Smart Grid* paper [22].

- In Chapter 5, a joint estimation framework for recovering missing data in urban systems is introduced. Specifically, the interaction between different components of the urban system gives rise to another type of correlation that is exploited in this setting. The theoretical conditions under which the joint recovery of two datasets requires fewer observations compared with the independent recovery of both datasets are characterized within a matrix completion framework. Moreover, when each dataset is sampled uniformly at random, lower bounds for the probability of recovery are provided for both the independent and the joint recovery of two datasets. The probability bounds are then used to optimize the sampling process in order to minimize the acquisition cost for the observations used in the recovery process. Numerical studies to benchmark the performance of different algorithms that exploit the additional type of correlation are also presented. The numerical results demonstrate that new algorithm proposed in Chapter 4 outperforms the matrix completion-based recovery in a joint recovery setting. Moreover, the novel approach is robust to different sampling regimes in the datasets combined.

# Chapter 2

# Overview of theory and algorithms for missing data recovery

This chapter presents a review of the state-of-the-art theoretical results and algorithms for addressing the missing data recovery problem.

In the following, Section 2.1 presents an overview of different approaches for addressing the missing data problem. Section 2.2 summarizes the key theoretical results in the matrix completion area and shows the main application domains for this type of methods. In addition, the main algorithmic approaches are presented. The conclusions are presented in Section 2.3 and a standard matrix completion-based algorithm is adopted for benchmarking purposes using real data collected in an urban system.

## 2.1 Missing data recovery methods

The recent advances in both sensing and communication technologies facilitated the deployment of large sensing networks across different components of the urban system [9]. The significant boost in the sensing capabilities of the urban system gives rise to new challenges for the data acquisition and data processing tools [15]. In particular, the large number of sensors deployed provides measurements over vast geographical areas and long periods of time which results in high-dimensional datasets. The measurements are used as inputs for data analysis tools, or modeling and optimization techniques which help the operator in the process of controlling and managing the system as well as in the decision making process [15]. However, timely and accurate observations are necessary for the data processing tools to achieve their goal. In practice, measurements are lost in the acquisition process, and therefore, the practical reach of the data processing techniques is limited. Consequently, the missing data problem needs to be addressed at the acquisition stage.

Assuming the data is organized in matrix form, one approach to deal with missing data is the listwise deletion in which the rows or the columns that contain missing values are deleted [24], [25] and [26]. This approach is easy to implement but leads to a decrease in the size of the dataset and loss of information. To avoid this scenario it is necessary to estimate the missing values based on the available observations. The methods for estimating missing data can be grouped in three main categories [27]: Imputation-based methods, Matrix-based methods, and Statistics-based methods. These are reviewed in more detail in the following sections.

### 2.1.1 Imputation-based methods

Imputation is the process of replacing a missing entry with a statistical prediction [28]. The main aim of this type of methods is to fill the gaps in the matrix without introducing a large amount of bias. Depending on the number of imputations required for each missing entry, this type of methods are further divided into two categories:

1. *Single imputation methods* replace the missing values with the mean of other values in the dataset. In [29] and [25] the mean is calculated using a fixed number of neighbor entries. The main advantage of this approach is that no information is removed from the dataset. On the other hand, the main disadvantage is that the output depends on the number of neighbor entries considered which creates bias in the dataset [24].

2. *Multiple imputation methods* generate $m$ datasets by estimating the missing values $m$ times. The final value is obtained by averaging the imputed values for the same data point in all the datasets generated [30], [31] and [32]. This type of methods aim to generate unbiased datasets and to provide confidence coefficient values to show the reliability of the estimated values [27]. Guidelines for selecting the number of datasets are proposed in [33] and [34]. The complexity that arises from processing multiple datasets is one of the main limitations of this type of methods. In addition, the multiple imputation-based estimation generates non-deterministic results [27].

### 2.1.2 Matrix-based methods

Matrix-based methods operate under the assumption that the entries contained in the data matrix are correlated, and consequently, the resulting data matrix is low rank or approximately low rank. In this context, the recovery of the missing entries of low rank matrices is feasible in a convex optimization context provided that a sufficient fraction of the entries is observed [35], [36], [37], and [38]. In [29] a mixture between imputation-based methods and matrix-based methods is used to replace missing

entries with the row average and then compute the most significant eigenvalues and the corresponding eigenvectors. The missing values are estimated using a linear regression of the obtained eigenvectors. A more computationally efficient method is proposed in [39]. Therein, the Singular Value Thresholding (SVT) algorithm obtains the matrix with minimum rank that fits the observations by minimizing the nuclear norm, i.e. the sum of the singular values of the matrix. Matrix completion approaches have been successfully applied in a wide range of practical settings. For instance, the performance of the matrix completion approaches in recovering missing data for an electricity distribution system is studied in [21], [19] and [22]. A review of the techniques used to estimate the state of the electricity distribution system is presented in [40]. In addition, low rank minimization tools are used for electricity price forecasting in [41].

### 2.1.3 Statistics-based methods

This type of methods exploit the statistical properties of the dataset in order to estimate the missing values. Specifically, a recovery method that exploits the hierarchical relationship between time series data to apply a locally weighted regression is proposed in [42]. In [43], a local least squares recovery method that estimates the missing data as a linear combination of similar genes is presented. The similar genes are selected based on the absolute value of the Pearson correlation coefficients [44]. For the cases in which not all the observations are correlated, in [45] the authors apply a locally weighted regression only on the locally correlated observations. The selection of the observations is done using the Least Absolute Shrinkage and Selection Operator (LASSO). However, due to the multiple steps required for selection and validation for each missing entry, this approach has a large computational cost. Moreover, there are no guidelines for choosing the optimal number of nearest neighbors [27]. Alternatively, the statistical properties of the voltage data collected from the electricity distribution grid are exploited in a Bayesian framework to recover missing data in [19]. The main disadvantage of this approach is that it requires prior knowledge about the second order statistics of the dataset. In addition, the performance of the recovery varies significantly when the prior knowledge is not accurate.

A more in depth description of the matrix-based methods for recovering missing data follows.

## 2.2 Matrix completion

Matrix completion is a relatively new research area that emerged in 2009 when Candès and Recht proposed in [35] a convex relaxation of the NP-hard problem of finding

the matrix with the minimum rank that fits the available observations. This type of recovery exploits the low rank structure typically found on datasets with correlated entries. Respectively, correlated datasets give rise to low rank or approximately low rank data matrices. The missing entries are recovered by computing the matrix with the minimum rank that fits the available observations. Therefore, the matrix completion recovery method is formulated as

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \text{rank}(\mathbf{X}) \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \end{aligned} \tag{2.1}$$

where $\mathbf{X}$ is the optimization variable, $\mathbf{M}$ is the matrix that needs to be estimated, $\Omega$ is the set containing positions of the available observations in the matrix $\mathbf{M}$, and $P_\Omega$ is the orthogonal projector into the subspace of matrices that vanish outside of $\Omega$. Unfortunately, the rank minimization problem in (2.1) is NP-hard and all the algorithms achieving the exact solution require time at least doubly exponential in the dimension of the matrix [46]. Favorably, in [35] it is shown that when the entries on $\Omega$ are sampled uniformly at random, the solution of the rank minimization problem in (2.1) is obtained with high probability by solving the following convex relaxation

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \end{aligned} \tag{2.2}$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm of the matrix $\mathbf{X}$. However, not any matrix can be recovered from a subset of the entries. To better understand the limits of the matrix completion framework, the main results in [35] are summarized below.

A matrix $\mathbf{M}$ of size $M \times N$ and rank $r$ is recovered with probability $1 - cN^{-3}\log N$ by solving the optimization problem in (2.2) when the following conditions are satisfied:

1. the matrix $\mathbf{M}$ is sampled from the random orthogonal model,
2. the locations of the missing entries are sampled uniformly at random,
3. the number of entries in $\mathbf{M}$ that are observed obeys

$$k \geq C\, N^{1.25}\, r \log N, \tag{2.3}$$

where $C$ and $c$ are constants and $N = \max\{M, N\}$.

To define the random orthogonal model, let us consider the singular value decomposition of the matrix $\mathbf{M}$

$$\mathbf{M} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T, \tag{2.4}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ for $j = \{1, 2, ..., r\}$ are the left and right singular vectors and $\sigma_j$ are the singular values. In this context, a generic low rank matrix is sampled from the random orthogonal model if the families $\{\mathbf{u}_j\}_{1\leq j\leq r}$ and $\{\mathbf{v}_j\}_{1\leq j\leq r}$ are sampled uniformly at random among all collections of $r$ orthonormal vectors independently of each other. Note that the results not only define the conditions in which there exists a low rank solution that is consistent with the available observation but it also shows that the convex relaxation in (2.2) achieves the same solution as the NP-hard problem in (2.1).

Since the applicability of this result is limited to the class of matrices that belong to the random orthogonal model, a more general result that covers a larger class of matrices $\mathbf{M}$ is also presented in [35]. A matrix $\mathbf{M}$ of size $M \times N$ with rank $r$ is recovered with probability at least $1 - cN^{-\beta}$ when the following conditions are satisfied:

1. the matrix $\mathbf{M}$ obeys A0 and A1,
2. the locations of the missing entries are sampled uniformly at random,
3. the number of entries in $\mathbf{M}$ that are observed obeys

$$k \geq C \max\{\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 N^{1/4}\}\, N\, r\, \beta \log N, \tag{2.5}$$

where $C$ and $c$ are constants, $\beta > 2$, $N = \max\{M, N\}$ and $\mu_0$ and $\mu_1$ are defined in the following.

Let $U$ be a subspace of $\mathbb{R}^N$ of dimension $r$ and $\mathbf{P}_U$ be the orthogonal projection onto $U$. Then the coherence of $U$ (vis-a-vis the standard basis $(\mathbf{e_i})$) is defined to be

$$\mu(U) \equiv \frac{N}{r} \max_{1\leq i\leq N} \|\mathbf{P}_U \mathbf{e_i}\|^2, \tag{2.6}$$

where $\mathbf{e_i}$ is a vector with 1 on the $i$-th position and zero in rest.

Let $\mathbf{USV}^T$ be the singular value decomposition of matrix $\mathbf{M}$. Where $\mathbf{S}$ is $r \times r$, $\mathbf{U}$ is $M \times r$, $\mathbf{V}$ is $N \times r$ and $r$ is the rank of $\mathbf{M}$. In this context, the two assumptions are

A0: there exists $\mu_0 > 0$ such that $\max\{\mu(\mathbf{U}), \mu(\mathbf{V})\} \leq \mu_0$.

A1: there exists $\mu_1 > 0$ such that $\mu_1 \equiv \frac{\max_{1\leq i\leq M, 1\leq j\leq N} (\mathbf{UV}^T{}_{i,j})}{\sqrt{r/(MN)}}$.

In other words, the original problem is equivalent to minimizing the number of nonvanishing singular values while the convex relaxation minimizes the sum of the singular values. The fact that the nuclear norm minimization problem produces a low rank solution is observed in [47], [48], and [49] but the first theoretical guarantee is provided in [35]. The bounds for the minimum number of entries required are subsequently improved in [36] and in [50]. Another significant contribution is

presented in [51] where the bound for the number of entries required is optimized up to a small numerical constant and a logarithmic factor. The main result in [51] follows.

A matrix $\mathbf{M}$ of size $M \times N$ with rank $r$ is recovered with probability at least $1 - 6\log(N)(M+N)^{2-2\beta} - N^{2-2\beta^{1/2}}$ by solving the optimization problem in (2.2) when the following conditions are satisfied:

1. the matrix $\mathbf{M}$ obeys A0 and A1,
2. the locations of the missing entries are sampled uniformly at random,
3. the number of entries in $\mathbf{M}$ that are observed obeys

$$k \geq 32 \max\{\mu_1^2, \mu_0\}\, r(M+N)\, \beta \log^2(2N), \tag{2.7}$$

where $\beta > 1$ and $M \leq N$.

The results above are of limited applicability in practical settings because it is assumed that the available observations are noise free. Matrix completion approaches to missing data recovery from noisy observations are studied in [38], [37], [52], [53], [54], [55], [56], [57], and [58]. Interestingly, Theorem 7 in [38] shows that the matrix completion framework is robust to noise. Specifically, when the matrix $\mathbf{M}$ satisfies the conditions for perfect recovery in the noiseless setting, the recovery is also feasible using a noisy subset of observations. Moreover, the recovery error is proportional to the noise level [38].

An information-theoretic characterization of the fundamental limits is provided in [59], where it is shown that low rank matrices are recovered from $k$ linear measurements when

$$k > (M + N - r)r. \tag{2.8}$$

The linear measurements are in the form of

$$\mathbf{y} = (\langle \mathbf{B}_1, \mathbf{M} \rangle, \ldots, \langle \mathbf{B}_k, \mathbf{M} \rangle)^T \in \mathbb{R}^k, \tag{2.9}$$

where $\mathbf{B}_i \in \mathbb{R}^{M \times N}$ denotes the measurement matrices with $i \in \{1, 2, \ldots, k\}$ and $\langle \cdot, \cdot \rangle$ denotes the trace inner product between matrices in $\mathbb{R}^{M \times N}$. One of the strengths of the result in [59] is that the linear measurements model allows for more flexibility in the sampling process compared to the uniform sampling model. In addition, the recovery limits are characterized for an ensemble of matrices described by their probability, and therefore, linking matrix completion to a Bayesian formulation.

Matrix completion has applications in various domains. A few examples are provided in [38] and include:

1. *Collaborative filtering*: to make predictions about the interests of an user based on information from many other users [60]. One example of such a problem is the Netflix recommendation system in which the objective is to make rating predictions about unseen movies [35].

2. *Global positioning*: to find the position in the Euclidean space from a local or partial set of pairwise distances. This type of problems arise in sensor networks [61], [62], and [63].

3. *System identification*: to fit a discrete-time linear time-invariant state-space model to a sequence of inputs and outputs [64], [49], [48], [65], and [66].

4. *Remote sensing*: to determine the direction of arrival of incident signals in a coherent ratio-frequency environment [67], and [38].

5. *Structure-from-motion*: to recover a scene and infer the camera motion from a sequence of images [68], and [69].

6. *Multiclass learning* in data analysis [70], [71], and [72].

The joint recovery problem for different datasets is addressed within a rank minimization framework. Specifically, matrix completion results are extended to the tensor case in [73] by defining the trace norm for tensors. Motivated by the matrix completion case, the tensor completion is formulated as a convex optimization problem for which several recovery algorithms are proposed in [73]. An alternative approach to tensor completion is presented in [74] using Riemannian optimization techniques. The singular value decomposition is also extended to the tensor case in [75] which leads to the development of a tensor nuclear norm penalized algorithm in [76]. A weighted tensor low-rank regularization approach that exploits spatial, spectral and temporal information for recovering images in remote sensing is presented in [77]. In addition to the tensor recovery problem, a collective matrix completion framework is proposed in [78] and [79] to exploit the correlation between different datasets.

Matrix completion algorithms that solve the rank minimization problem can be divided in two categories:

1. **Convex optimization-based algorithms**: find the matrix with the minimum rank by solving the nuclear norm minimization problem proposed in [35]. Note that the convex relaxation only works under certain sampling conditions. The computational time is polynomial in the matrix dimension. Also, this type of algorithms can be further grouped into two main categories depending on the complexity and the computational costs:

(a) *Interior-point methods* are implemented using general semidefinite programming solvers such as SDPT3 [80] or SeDuMI [81] provide high numerical accuracy. However, the trade-off is a high computational cost and high memory requirements. In fact, SDPT3 can only handle matrices with sizes up to one hundred. In [64] the structure of the problem is exploited to reduce the memory requirements and increase the matrix size up to 350.

(b) *First-order methods* have a lower computational cost per iteration and are more flexible for specific structures of semidefinite programs [82]. The list of such algorithms includes: accelerated proximal gradient (APG) descent [83] which is based on the fast iterative shrinkage-thresholding algorithm (FISTA) for matrix completion [84], conditional gradient with enhancement and truncation (CoGENT) [85], alternating descent conditional gradient (ADCG) [86], augmented Lagrangian multiplier methods [87], Frank-Wolfe [88], [89], [90], and the singular value thresholding (SVT) [39].

2. **Nonconvex optimization-based algorithms**: solve the original nonconvex rank minimization problem at a significantly reduced computational cost compared to that of the convex optimization-based methods. The key idea for this type of methods is to use the Burer-Monteiro factorization [91] that enforces the low-rank constraint through a low rank factorization of the objective matrix. Consequently, the objective function becomes a function of the two factor variables. In [82], this type of methods are divided into three classes of iterative schemes:

   (a) *Projected gradient descent methods* run gradient descent directly of the objective function with respect to the factor variables [91], [92] and [93].

   (b) *Alternating minimization methods* optimize the cost function alternatively over one factor while fixing the other [94], [95].

   (c) *Singular value projection methods* perform a gradient descent step of the cost function on the $M \times N$ matrix space and then project back to the factor space using the singular value decomposition [96], [97], [98].

## 2.3 Conclusions

This chapter presented an overview of different methods used for estimating missing data. The main approaches are categorized based on their complexity and the underlying assumptions required for the recovery process. A relatively new research area, i.e., matrix completion, is explored in more detail to verify if the main assumptions

are satisfied in urban high-dimensional datasets. The main application domains for matrix completion-based methods are presented and an overview of the different algorithmic approaches is included. In this context, one of the algorithms is selected for numerical evaluation using real data from an urban system. A low computational cost is key for the applicability of the algorithm in the high-dimensional urban datasets.

The SVT algorithm proposed in [39] is a first-order method that solves the nuclear norm minimization problem. In addition, the structure of the problem is exploited to further reduce the computational cost and the memory requirements. To that end, capitalizing on the sparsity and the low rank properties of the matrices computed at each iteration allows the SVT algorithm to recover matrices with close to a billion entries in 17 minutes on a computer with 1.86GHz and 3 GB of memory [39]. In addition, theoretical guarantees for the convergence of the algorithm are provided in [39]. For this reasons, the SVT is adopted for benchmarking purposes in this thesis. However, the main limitation of the SVT algorithm is that it requires parameter tuning and there are no guidelines for optimizing the value for different recovery scenarios. A more in depth description of the SVT approach is presented in Section 3.3.

Robust recovery methods are required to guarantee the availability and reliability of monitoring data in urban settings. Therefore, the limitation of the SVT algorithm is addressed to increase the resilience of the monitoring process to missing data. For that reason, when the missing data cannot be recovered because the number of available entries is not sufficient, alternative strategies need to be considered. Since this problem is not currently considered in the literature, a joint estimation framework for multiple datasets is presented. The proposed estimation process leverages the correlation between different data sources to jointly recover the missing data from datasets with insufficient observations. However, the joint estimation of two or more datasets is not always advantageous. To identify the conditions that yield beneficial joint recovery, the fundamental limits of the joint recovery setting are derived in order to provide guidelines for combining different types of data.

In the following chapter, the numerical performance of the SVT algorithm is benchmarked against other recovery strategies and the optimal performance theoretically attainable. The comparison with the fundamental limit determines the scenarios in which the performance of the state of the art algorithms is largely suboptimal and demonstrates the need for a novel recovery method that addresses the limitation of the SVT approach.

# Chapter 3

# A performance evaluation framework for missing data recovery methods

## 3.1 Introduction

This chapter provides a framework for analyzing and comparing the performance of different estimation methods for the recovery of missing data in high-dimensional datasets. In addition, the fundamental limits of missing data recovery from noisy observations are characterized using information theoretic results. In particular, the proposed framework is demonstrated using data from a low voltage distribution grid. This dataset is an example of data collected in the urban setting, is high-dimensional and highly correlated with other areas of human activity and urban processes. The availability of real data enables benchmarking different estimation methods in a practical setting.

In the following, Section 3.2 formulates the missing data recovery problem in mathematical terms and introduces a model for the real data collected from an electricity distribution system. Section 3.3 presents the SVT algorithm that is used for benchmarking purposes. In Section 3.4, the Linear Minimum Mean Squared Error (LMMSE) estimator is introduced as it provides optimal recovery for the model presented in Section 3.2. In addition, the fundamental limits of the data recovery problem are characterized by defining the optimal performance theoretically attainable by any estimator. A numerical comparison using real data is presented in Section 3.5 and the limitations of the state-of-the-art estimation methods are identified. Finally, the conclusions are presented in Section 3.6.

## 3.2 System model

This section defines the missing data recovery problem in a mathematical framework. Consider a low voltage (LV) distribution system with $N$ feeders. Each feeder includes a sensing unit that measures the electrical magnitudes of operational interest at predetermined time instants. These measurements include phase active power, phase reactive power and phase voltage and support the operator in controlling, monitoring, and managing the network. In practice, due to telemetry errors, the acquisition process provides the operator with a noisy and incomplete set of measurements. For that reason, the operator needs to recover the missing LV data using the available observations.

### 3.2.1 Data Source Model

For a given electrical magnitude $s$, let $m_{i,j}^{(s)}$ be the corresponding value at feeder $i \in \{1, 2, ..., N\}$ at time $j \in \{1, 2, ..., M\}$. The matrix with measurements for magnitude $s$, denoted by $\mathbf{M}^{(s)} \in \mathbb{R}^{M \times N}$, contains the aggregated measurement vectors from all feeders

$$\mathbf{M}^{(s)} \triangleq [\mathbf{m}_1^{(s)}, \mathbf{m}_2^{(s)}, ..., \mathbf{m}_N^{(s)}], \tag{3.1}$$

where the measurement vectors are given by

$$\mathbf{m}_i^{(s)} = [m_{i,1}^{(s)}, m_{i,2}^{(s)}, ..., m_{i,M}^{(s)}]^T, \tag{3.2}$$

and $\mathbf{m}_i^{(s)} \in \mathbb{R}^M$. Without loss of generality the analysis is carried out for a particular electrical magnitude i.e., phase voltage, and therefore, the index $s$ is dropped. The resulting data matrix $\mathbf{M}$ contains the phase voltage measurements at time instants $1, 2, ..., M$ for all $N$ feeders.

In the following, actual LV data is used to model the statistical structure of the data generated in a LV electricity distribution system. The actual LV dataset under consideration contains values from 200 residential secondary substations across the North West of England collected from June 2013 to January 2014. The data collection is part of the “Low Voltage Network Solutions" project run by Electricity North West Limited [99]. Each substation creates a daily file with measurements including minimum, maximum and mean values for the voltage, current and power levels on all three phases (A, B, C). The measurements are collected every minute.

Using the dataset shared by ENWL, a matrix $\mathbf{M}$ is constructed such that $M = N = 500$, where $M$ is the number of minutes and $N$ corresponds to the number combinations between feeders and days. Note that the feeders and the dates are randomly selected.

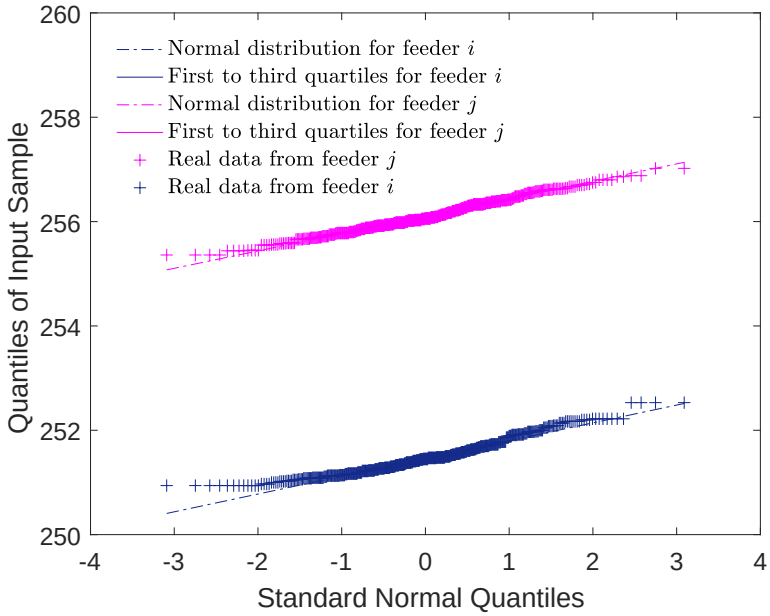



Figure 3.1. Quantile-Quantile plot for the distribution of the phase A mean voltage data provided by ENWL versus a Normal distribution.

In this context, each entry of the matrix $\mathbf{M}$ is a value that describes the state of the grid in phase A for a given feeder and day when averaged over a one minute period. The data collection and storage is managed by ENWL and no pre-processing of the voltage measurements is required before the recovery process. The test matrix $\mathbf{M}$ is obtained by randomly extracting a complete set of measurements from the data files provided by ENWL. Hence, there are no missing entries in the test matrix $\mathbf{M}$. The rationale for this is to randomly discard some of the entries and then benchmark the performance of different recovery strategies against the actual measurements.

The analysis that follows pertains to the extracted data matrix $\mathbf{M}$ with phase A mean voltage measurements but the results apply to the data matrices containing the observations from phases B and C, and therefore, the analysis is provided for phase A without loss of generality.

The comparison depicted in Figure 3.1, between two columns of the matrix $\mathbf{M}$ and the corresponding normal distributions highlights that the distribution of the phase A mean voltage data is approximately Gaussian with a slight deviation on the tails. Hence, the distribution of the real data is well approximated by a Gaussian distribution.

Moreover, the sample covariance matrix of the matrix containing phase A mean voltage measurements is presented in Figure 3.2. In is worth noting that the covariance matrix exhibits a structure that is approximately Toeplitz. Because the voltage data is correlated, the covariance matrix displays a high correlation across feeders and time instants. The Toeplitz model resembles a physical temporal correlation where

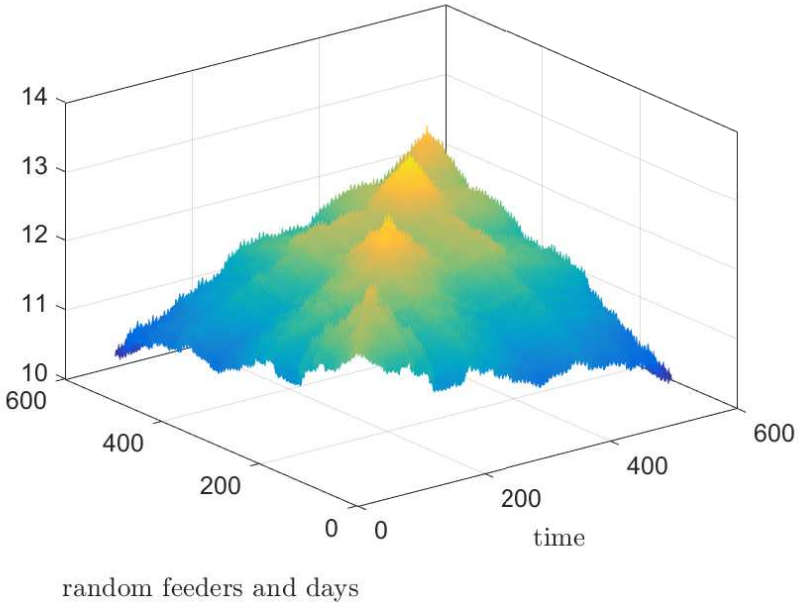



Figure 3.2. Sample covariance matrix of the phase A mean voltage data provided by ENWL.

the correlation decreases as the temporal distance increases. This implies that the correlation between two voltage measurements in the same feeder depends on their separation in time.

Assuming the data is stationary, the LV data is modelled as a multivariate Gaussian distribution with a Toeplitz covariance matrix. The rationale for this assumption is backed by the Gaussianity of the data in different feeders observed in Figure 3.1 and the structure of the sample covariance matrix depicted in Figure 3.2. An analysis of the distribution and sample covariance matrix of the same voltage measurements is presented in [19]. Therein, it is shown that voltage measurements can be modelled as a multivariate Gaussian random process, more specifically for all $i \in \{1, 2, ..., N\}$, it is assumed that

$$\mathbf{m}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.3}$$

and $\mathbf{m}_i$ is a sequence of independent and identically distributed random variables. Consequently, $\mathbf{M}$ is a realization of the random process describing the value of the voltage measurements across the grid.

Another consequence of the correlation across feeders and time instants displayed in the covariance matrix, is that the singular value decomposition of the matrix $\mathbf{M}$ has a large condition number [100]. In other words, the ratio of the largest to smallest singular value is large. Figure 3.3 shows the singular values of the matrix $\mathbf{M}$ in decreasing order. Remarkably, there are a few singular values that concentrate most

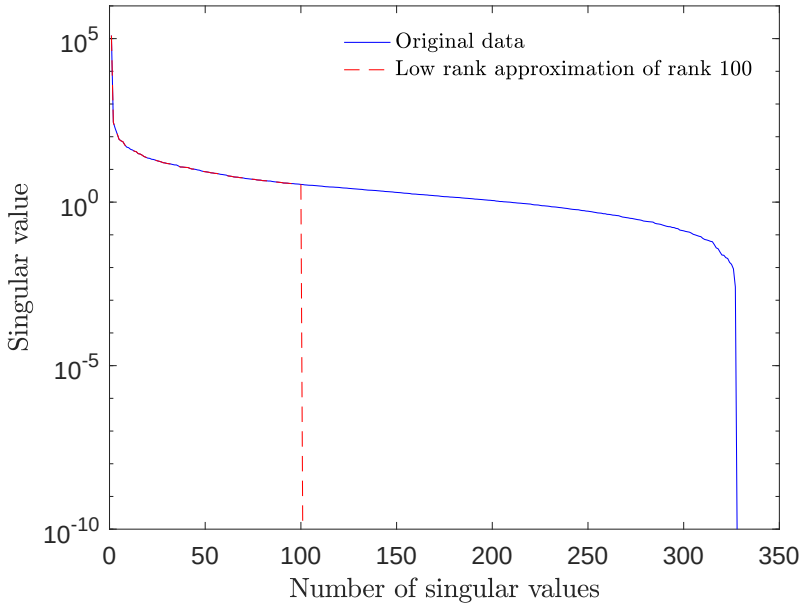



Figure 3.3. Singular values of the matrix $\mathbf{M}$ containing actual voltage measurements, when $M = N = 500$, compared to a low-rank approximation of the same matrix when the rank is one hundred.

of the norm of the matrix. Precisely, the first five singular values in decreasing order concentrate 98.78% of the matrix nuclear norm while the first thirty singular values concentrate 99.4% of the matrix nuclear norm. That being the case, the truncated low-rank approximation of the matrix obtained by setting the $\text{rank}(\mathbf{M}) - L$ smallest singular values to zero does not introduce a significant approximation error for some values of $L$. Therefore, the matrix $\mathbf{M}$ can be modelled as approximately low rank, see [35], [36], and [38]. In this context, the problem of estimating the matrix $\mathbf{M}$ can be posed as a rank minimization problem. However, in this case, the performance of the rank minimization based algorithm is lower bounded by the low rank approximation error. This is caused by the fact that the matrix $\mathbf{M}$ is approximately low rank and therefore the recovery error between $\mathbf{M}$ and any low rank approximation of $\mathbf{M}$ produced by a rank minimization based algorithm is bounded by the NMSE depicted in Figure 3.4.

The voltage measurements aggregated in the matrix $\mathbf{M}$ describe the state of the LV grid during $M$ time instants. However, in practice, the operator does not have access to the actual measurements describing the grid. Instead, part of the observations collected by the monitoring system are missing and the ones that are available are corrupted by noise.

Figure 3.5 describes the distribution system monitoring model. In this setting, the voltage measurements describing the state of the system are modelled as a random process that generates a realization $\mathbf{M} \in \mathbb{R}^{M \times N}$ every $M$ time instants. The state

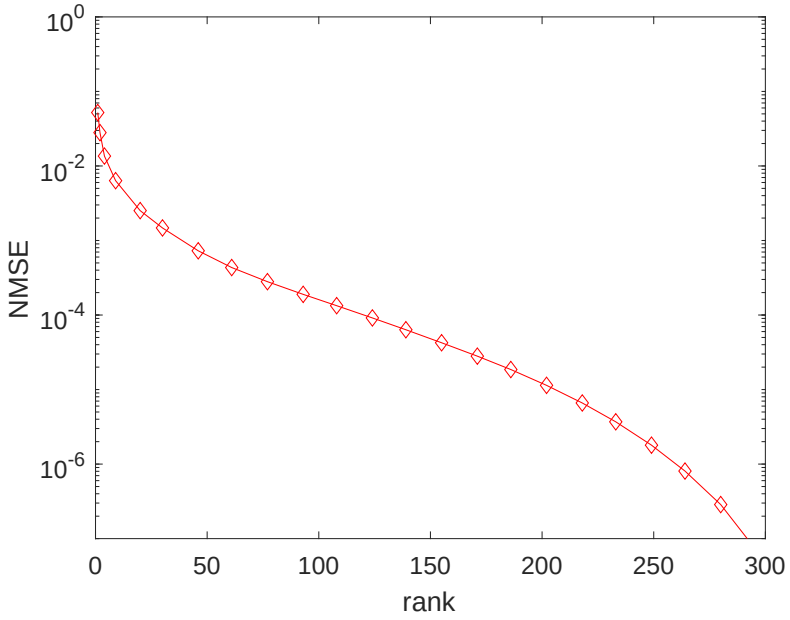



Figure 3.4. NMSE between $\mathbf{M}$ and the low rank approximation for different rank values.

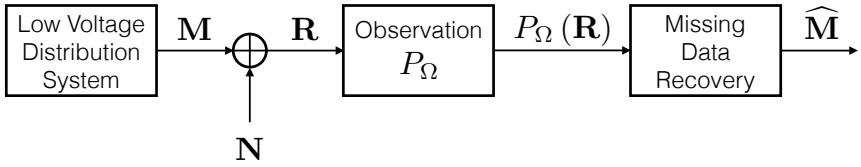



Figure 3.5. Block diagram describing the system model.

of the grid is fully described by the entries of the matrix $\mathbf{M}$. It is assumed that data is lost in the acquisition process so that only a subset of the complete set of observations is available. Moreover, the observations that are available are corrupted by noise originating in the sensing infrastructure from different types of noise sources such as thermal or electromagnetic. In this context, the aim of the estimation process is twofold: to denoise the available observations and to recover the missing entries.

### 3.2.2 Acquisition

The voltage measurements are assumed to be corrupted by additive white Gaussian noise (AWGN) such that the resulting observations are given by

$$\mathbf{R} = \mathbf{M} + \mathbf{N}, \tag{3.4}$$

where

$$(\mathbf{N})_{i,j} \sim \mathcal{N}(0, \sigma^2_{\mathbf{N}}), \tag{3.5}$$

where $(\mathbf{N})_{i,j}$ denotes the entry in row $i$ and column $j$ of the matrix $\mathbf{N}$ and $i \in \{1, 2, ..., M\}$ and $j \in \{1, 2, ..., N\}$. Moreover, it is also assumed that only a fraction of the complete set of measurements (entries in $\mathbf{R}$) are communicated to the operator. Denote by $\Omega$ the subset of observed entries, i.e., $\Omega \subseteq \{1, 2, ...M\} \times \{1, 2., ..., N\}$. By definition it follows that $\Omega$ is given by

$$\Omega \triangleq \{(i, j) : (\mathbf{R})_{i,j} \text{ is observed}\}. \tag{3.6}$$

Formally, the acquisition process is modelled by the function $f : \mathbb{R}^{M \times N} \to \mathbb{R}^{|\Omega|}$ with $f(\mathbf{M}) = P_\Omega(\mathbf{R})$ where

$$P_\Omega(\mathbf{R}) = (\mathbf{R})_\Omega, \tag{3.7}$$

and $|\Omega|$ denotes the cardinality of $\Omega$. The observations given by (3.7) describe all the data that is available to the operator for estimation purposes and therefore, the recovery of the missing data is performed from the observations $P_\Omega(\mathbf{R})$.

### 3.2.3 Estimation

The estimation process of the complete matrix of measurements based on the available observations is modelled by the function $g : \mathbb{R}^{|\Omega|} \to \mathbb{R}^{M \times N}$. The estimate $\widehat{\mathbf{M}} = g(f(\mathbf{M}))$ is obtained by solving an optimization problem based on an optimality criterion. For this work, the optimality criterion is the normalized mean square error (NMSE) given by

$$\text{NMSE}\,(\mathbf{M}; g) = \frac{\mathbb{E}\left[\|\mathbf{M} - g(f(\mathbf{M}))\|_F^2\right]}{\|\mathbf{M}\|_F^2}, \tag{3.8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. It is worth noting that the NMSE is proportional to the mean square error (MSE) for a fixed matrix $\mathbf{M}$ as it satisfies

$$\text{MSE}\,(\mathbf{M}; g) = \text{NMSE}\,(\mathbf{M}; g)\,\frac{\|\mathbf{M}\|_F^2}{MN}. \tag{3.9}$$

To summarize, the actual state of the LV distribution system is described by the set of measurements aggregated in the matrix $\mathbf{M}$. However, in the acquisition process some of the measurements are lost and the observations that are available are corrupted by AWGN. The estimation step produces an estimate $\widehat{\mathbf{M}}$ based on the noisy subset of entries that is available. The performance of the estimation procedure is evaluated based on the NMSE.

## 3.3 Singular Value Thresholding

SVT is a matrix completion-based algorithm [39] which produces a sequence of matrices $\mathbf{X}^{(k)}$ that converges to the unique solution of the following optimization

problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \tau\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X}\|_F^2 \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \end{aligned} \tag{3.10}$$

where $\|\mathbf{X}\|_*$ is the nuclear norm of the matrix $\mathbf{X}$ and $(k)$ denotes the iteration index. Note that when $\tau \to \infty$, the optimization problem in (3.10) converges to the nuclear norm minimization problem proposed in [35]

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}). \end{aligned} \tag{3.11}$$

For large values of $\tau$, the SVT algorithm provides the solution to the nuclear norm minimization problem.

The main steps of the SVT algorithm are:

$$\begin{cases} \mathbf{X}^{(k)} = D_\tau(\mathbf{Y}^{(k-1)}), \\ \mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_s\Big(P_\Omega(\mathbf{M}) - P_\Omega(\mathbf{X}^{(k)})\Big), \end{cases} \tag{3.12}$$

where $\mathbf{Y}^{(0)} = \mathbf{0}$, $\delta_s$ is the step size that obeys $0 < \delta_s < 2$, and the soft-thresholding operator, $D_\tau$, applies a soft-thresholding rule to the singular values of $\mathbf{Y}^{(k-1)}$, shrinking these towards zero. For a matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$ of rank $r$ with singular value decomposition given by

$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad \mathbf{S} = \text{diag}(\{\sigma_i(\mathbf{Y})\}_{1\le i\le r}), \tag{3.13}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of size $M \times r$ and $N \times r$, respectively, and $\sigma_i(\mathbf{Y})$ are the singular values of the matrix $\mathbf{Y}$, the soft-thresholding operator is defined as

$$D_\tau(\mathbf{Y}) \triangleq \mathbf{U}D_\tau(\mathbf{S})\mathbf{V}^T, \text{ with } D_\tau(\mathbf{S}) = \text{diag}(\{(\sigma_i(\mathbf{Y}) - \tau)_+\}), \tag{3.14}$$

where $t_+ = \max\{0, t\}$. Interestingly, the choice of $\tau$ is important to guarantee a successful recovery, since large values guarantee a low-rank matrix estimate but for values larger than $\max_i(\sigma_i(\mathbf{Y}))$ all the singular values vanish. In [39], the proposed threshold is $\tau = 5N$. However, simulation results presented in [19] show that $\tau = 5N$ gives suboptimal performance when the number of missing entries is large. The choice of $\tau$ governs the performance trade-off between the high and low sampling regimes. Large values of $\tau$ yield a good performance when a small number of observations is available. Conversely, smaller values of $\tau$ yield a good performance when a large number of observations is available. Compared to alternatives like SeDuMi [81] or SDPT3 [80], SVT features a lower computational cost per iteration. This is achieved

by exploiting the sparsity of $\mathbf{Y}^{(k)}$ and the low-rank property of $\mathbf{X}^{(k)}$ to reduce storage requirements.

## 3.4 Performance limits

This section introduces two comparison terms for the performance of the matrix completion based recovery. First, the Linear Minimum Mean Squared Error estimator is introduced because it guarantees optimal recovery in this framework when the second order statistics are available. Precisely, the LMMSE estimator achieves the optimal performance in the recovery of missing data when:

1. the data is generated by a multivariate Gaussian source,
2. the optimality criteria is the MSE,
3. a subset of the entries is observed,
4. prior knowledge in the form of second order statistics is available.

For the estimation problem formulated in Section 3.2, the first three requirements for the optimality of the LMMSE estimator are satisfied. However, in practical applications, the actual second order statistics are not available. Instead, postulated statistics which are mismatched with respect to the actual ones, are available for estimation purposes. In this context, the performance degradation when mismatched statistics are available is studied in the Numerical results section.

The second benchmarking criteria presented in this section is the Optimal Performance Theoretically Attainable (OPTA) by any estimator when the data follows a multivariate Gaussian distribution. Precisely, the Rate-Distortion function which determines the achievable distortion for a given number of observations. This is an information theoretic limit and the optimal distortion is calculated based on the capacity of the AWGN channel that models the data acquisition process.

### 3.4.1 Linear Minimum Mean Squared Error estimator

MMSE estimation achieves the optimal performance in the recovery of missing data for a given set of observations $\Omega$ when the data is generated by a multivariate Gaussian source and the optimality criteria is the MSE. However, this estimation procedure relies on access to second order statistics of the state variables. The optimal estimate of the missing data given by the MMSE estimate is

$$\widehat{\mathbf{M}}_{\text{MMSE}} = \mathbb{E}[\mathbf{M}|f(\mathbf{M}), \mathbf{\Sigma}], \tag{3.15}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ is the covariance matrix defined in (3.3). In this framework, the sampling function $f$ is linear and is characterized by the set of matrices $\mathbf{A}_i$ such that

$$\mathbf{A}_i \mathbf{m}_i = P_\Omega(\mathbf{m}_i), \tag{3.16}$$

where $\mathbf{m}_i$ is the column $i$ of the matrix $\mathbf{M}$, with $i \in \{1, 2, ..., N\}$. The set of matrices $\mathbf{A}_i$ define a linear sampling operator that achieves the same set of observed entries as $P_\Omega$. Consequently, the MMSE estimation is identical with the Linear MMSE estimation.

In contrast with the matrix completion-based recovery, where all the matrix entries are estimated simultaneously, the LMMSE estimator operates for each column individually. For a matrix $\mathbf{M} \in \mathbb{R}^{M \times N}$, each of the $N$ columns is estimated independently.

The available entries from the column $i$ of the matrix $\mathbf{R}$ are given by

$$P_\Omega(\mathbf{r}_i) = \mathbf{A}_i(\mathbf{m}_i + \mathbf{n}_i), \tag{3.17}$$

where $\mathbf{n}_i$ is the column $i$ of the matrix $\mathbf{N}$ and $i \in \{1, 2, ..., N\}$. Therefore, the LMMSE estimate for each column of the matrix is given by

$$\widehat{\mathbf{m}}_i = \boldsymbol{\mu} + \mathbf{\Gamma}_i(P_\Omega(\mathbf{r}_i) - \mathbf{A}_i \boldsymbol{\mu}), \tag{3.18}$$

where $\boldsymbol{\mu}$ is presented in (3.3) and

$$\mathbf{\Gamma}_i = \mathbf{\Sigma}\mathbf{A}_i^T(\mathbf{A}_i \mathbf{\Sigma} \mathbf{A}_i^T + \sigma^2 \mathbf{I})^{-1}, \tag{3.19}$$

where $\mathbf{\Sigma}$ is the covariance matrix presented in (3.3) and $i \in \{1, 2, ..., N\}$. In order to compare the performance of the LMMSE estimation with the matrix completion based approach, the estimates $\mathbf{m}_i$ for $i \in \{1, 2, ..., N\}$ are aggregated in the matrix

$$\widehat{\mathbf{M}}_{\text{LMMSE}} = [\widehat{\mathbf{m}}_1, \widehat{\mathbf{m}}_2, ..., \widehat{\mathbf{m}}_N]. \tag{3.20}$$

Note that the estimates $\widehat{\mathbf{m}}_i$ defined in (3.18) depend on the covariance matrix $\mathbf{\Sigma}$. Therefore, obtaining the optimal estimate $\widehat{\mathbf{M}}_{\text{LMMSE}}$ requires knowledge of the probability distribution describing the data. If the data follows a Gaussian distribution it is equivalent to the knowledge of the second order moments, i.e. the covariance matrix $\mathbf{\Sigma}$ which needs to be known prior to the estimation process. In practice, the operator relies on postulated statistics that typically do not match the actual statistics. Consequently, the accuracy of the estimate is a function of the difference between the real and the postulated statistics.

### 3.4.2 Optimal Performance Theoretically Attainable

In order to assess the performance of the missing data recovery techniques in absolute terms, this section introduces the optimal performance theoretically attainable by an estimator $g$ when the data follows a multivariate Gaussian distribution.

Before introducing the fundamental limit for recovery of missing data, some key information theoretic [101] definitions that are necessary for the understanding of the Rate-Distortion [102] result are presented.

Let $X$ be a random variable with alphabet $\mathcal{X}$ and cumulative distribution function $F(x) = \Pr[X \leq x]$, $x \in \mathcal{X}$. When $F(x)$ is continuous, the random variable $X$ is also said to be continuous. Moreover, if the derivative of $F(x)$ is defined with $f(x) = F'(x)$, and $\int_{-\infty}^{\infty} f(x) = 1$, then $f(x)$ is called the probability density function for $X$. The set where $f(x) > 0$ is called the support set of $X$.

**Definition 1.** *The differential entropy $h(X)$ of the continuous random variable $X$ with probability density function $f(x)$ is defined by*

$$h(X) = -\int_{S_X} f(x) \log f(x)\, dx, \tag{3.21}$$

*where $S_X$ is the support set of the random variable $X$.*

Note that the entropy is a measure of the average uncertainty in the random variable, and therefore, is equivalent to the average number of bits required to describe it. The following definitions extend the notion of differential entropy to more than one random variable.

**Definition 2.** *The differential entropy of the vector of random variables $X_1$, $X_2$, ..., $X_n$ with probability density function $f(\mathbf{x})$ is given by*

$$h(X_1,\, X_2, ...,\, X_n) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}. \tag{3.22}$$

**Definition 3.** *The conditional differential entropy $h(X|Y)$ of a pair of random variables with joint probability density function $f(x,\, y)$ and conditional probability density function $f(x|y)$ is given by*

$$h(X|Y) = -\int f(x,\, y) \log f(x|y)\, dx\, dy. \tag{3.23}$$

The conditional entropy measures the amount of information required to describe the outcome of the random variable $X$ given that the value of random variable $Y$ is available in the recovery process.

The following definition introduces mutual information which measures the amount of information that one random variable contains about another random variable.

**Definition 4.** *Given two random variables $X$ and $Y$ with probability density functions $f_X(x)$ and $f_Y(y)$, respectively, and joint probability density function $f(x, y)$ the mutual information $I(X;Y)$ between the two random variables is defined by*

$$I(X;Y) = \int f(x,\, y) \log \frac{f(x,\, y)}{f_X(x)\, f_Y(y)}\, dx\, dy. \tag{3.24}$$

In other words, the mutual information measures the reduction in the uncertainty of the random variable $X$ given that the value of random variable $Y$ is available in the recovery process.

The Rate-Distortion function [102] of an i.i.d. source with probability density function $f(x)$ and squared-error distortion function $d(x, \hat{x})$ is given by

$$R(D) = \min_{f(\hat{x}|x):\mathbb{E}[\hat{X}-X]\leq D} I(X;\hat{X}), \tag{3.25}$$

where the distortion $d(x, \hat{x})$ of a symbol $x \in \mathcal{X}$ is a measure of the cost of replacing the symbol $x$ with $\hat{x}$, $R$ is the rate and $D$ is the distortion limit. Based on equation (3.25), the Rate-Distortion determines the minimum number of bits per symbol that should be communicated over a channel such that the message can be reconstructed with a distortion upper bounded by $D$.

In the electricity distribution setting described above, the measurements are corrupted by additive white Gaussian noise which determines the finite rate at which information about the state of the grid is obtained from the observations.

Consider a Gaussian channel with input $X_i$ and output $Y_i$ given by

$$Y_i = X_i + Z_i, \tag{3.26}$$

where

$$Z_i \sim \mathcal{N}(0, \sigma_Y^2). \tag{3.27}$$

Then for a power constraint $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$, the capacity of the Gaussian channel with bandwidth $W$ is given by [102]

$$C = W \log\left(1 + \frac{P}{\sigma_Y^2}\right), \tag{3.28}$$

where the bandwidth is measured in hertz and the capacity in measured in bits per second. Consequently, in the electricity distribution setting, the optimal performance is bounded by the capacity of the AWGN channel that models the acquisition process, i.e.,

$$R(D) < C, \tag{3.29}$$

where $R(D)$ is the rate at which the source needs to be observed to achieve distortion $D$. In view of this, the OPTA for a multivariate Gaussian source is given by

$$R(D) \leq \frac{|\Omega|}{2MN}\log(1+\mathsf{snr}), \tag{3.30}$$

where the signal to noise ratio, denoted by $\mathsf{snr}$, is defined as

$$\mathsf{snr} \triangleq \frac{\frac{1}{M}\mathrm{Tr}(\mathbf{\Sigma})}{\sigma^2_{\mathbf{N}}}, \tag{3.31}$$

where $\sigma^2_{\mathbf{N}}$ is the variance of the white Gaussian noise, defined in (3.5). The Rate-Distortion function of a multivariate Gaussian process is computed using the following parametric equations [103]

$$\begin{cases} R(\theta) &= \frac{1}{M}\sum_{i=0}^{M-1}\max\{0, \frac{1}{2}\log\frac{\lambda_i}{\theta}\} \\ D(\theta) &= \frac{1}{M}\sum_{i=0}^{M-1}\min\{\theta, \lambda_i\}, \end{cases} \tag{3.32}$$

where $R$ is the source rate in bits/symbol, $D$ is the mean square error distortion per entry, $\lambda_i$ is the $i$ th largest eigenvalue of $\mathbf{\Sigma}$, and $\theta$ is a parameter. The NMSE theoretically attainable, NMSE($\mathbf{M}$; OPTA), is given by

$$\mathrm{NMSE}(\mathbf{M};\mathrm{OPTA}) = D\frac{MN}{\|\mathbf{M}\|_F^2}. \tag{3.33}$$

## 3.5 Numerical results

This section presents a comparison for the missing data recovery performance of the LMMSE estimator and the SVT algorithm, and the theoretical limit, OPTA, using real data collected from a LV grid. To this end, a data matrix $\mathbf{M}$ is constructed using a complete set of voltage measurements from a LV distribution system. An analysis of the properties of such a matrix is presented in Section 3.2. The second step is to sample the matrix in order to decide which set of entries is available and which is missing. The missing entries are then estimated using both recovery methods and the performance is benchmarked against the theoretical limit given by the OPTA.

The test matrix, $\mathbf{M}$, is a square matrix of size 500, i.e. $M = N = 500$, and contains voltage measurements covering the state of the grid for a period of 2 hours. Each column is a vector that contains measurements describing the state of the grid on a different day and for a different feeder. The specific days and feeder to be included in the data matrix $\mathbf{M}$ are selected at random. However, selection criteria based on the geographic location or the past energy consumption behavior should increase the correlation in the matrix and therefore lead to better recovery using the

matrix completion based approach. The data selection problem is not addressed in this work.

For a given matrix $\mathbf{M}$, the variance of the AWGN is determined based on the SNR

$$\text{SNR} \triangleq 10 \log_{10} \mathsf{snr}, \tag{3.34}$$

where $\mathsf{snr}$ is defined in (3.31).

The resulting data matrix is sampled uniformly at random in order to decide which set of entries is available and which is estimated. The sampling probability is given by

$$\Pr[(i,j) \in \Omega] = \frac{1}{MN} \mathbb{E}[|\Omega|]. \tag{3.35}$$

The sampling probability for each entry of the matrix is a function of the expected number of available entries. Let $\gamma$ be the expected value of the proportion of missing entries for the matrix $\mathbf{M}$, that is:

$$\gamma \triangleq 1 - \frac{1}{MN} |\Omega|. \tag{3.36}$$

In this framework, the different sampling regimes are evaluated by changing the value of $\gamma$. For small values of $\gamma$ the number of available entries is large while for large values of $\gamma$ most of the entries are missing and need to be estimated. The random nature of the sampling process might result in different recovery performance for the same value of $\gamma$. In view of this, numerical simulations evaluate the recovery performance on multiple realizations of $\Omega$ for the same proportion of missing entries.

The performance of the two estimation methods is compared in terms of the NMSE given by

$$\text{NMSE}(\mathbf{M}; \text{SVT}) = \frac{\|\mathbf{M} - \widehat{\mathbf{M}}_{\text{SVT}}\|_F^2}{\|\mathbf{M}\|_F^2}, \tag{3.37}$$

where $\widehat{\mathbf{M}}_{\text{SVT}}$ is the SVT estimate of $\mathbf{M}$ based on $P_\Omega(\mathbf{R})$, and

$$\text{NMSE}(\mathbf{M}; \text{LMMSE}) = \frac{\|\mathbf{M} - \widehat{\mathbf{M}}_{\text{LMMSE}}\|_F^2}{\|\mathbf{M}\|_F^2}. \tag{3.38}$$

where $\widehat{\mathbf{M}}_{\text{LMMSE}}$ is the output of the LMMSE estimation. Moreover, the numerical results below show an average NMSE that is calculated over 100 realizations of $\Omega$ for each value of $\gamma$. The rationale for this is to avoid the situation in which a particular sampling pattern benefits one of the algorithms in terms of the recovery performance. The average NMSE describes the expected recovery performance for each estimation method.

In the following, the performance of the SVT algorithm is compared with the LMMSE estimation and the OPTA when SNR = 20 dB. However, it is noted in [19]

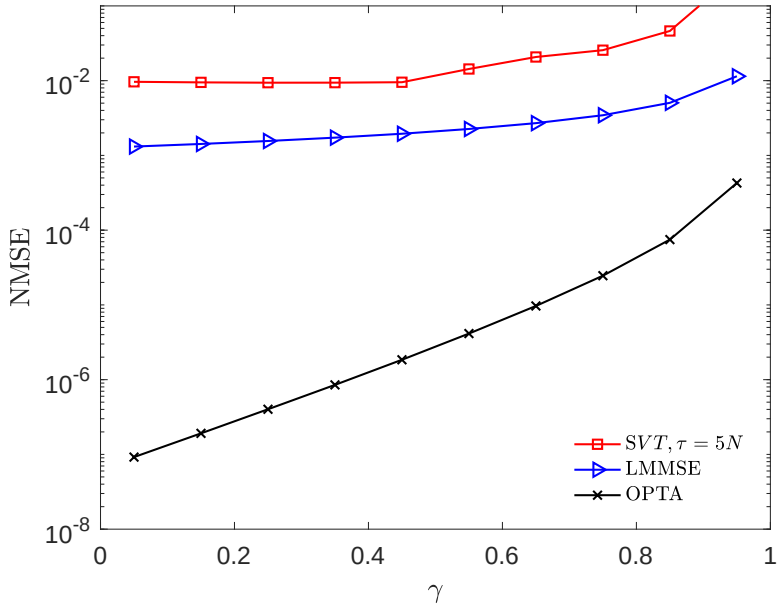



Figure 3.6. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT, LMMSE estimation, and the OPTA, when SNR = 20 dB.

that the performance of SVT recovery depends on the choice for the threshold $\tau$. In view of this, a performance comparison for different values of $\tau$ is also included. On the other hand, the performance of the LMMSE estimator relies on access to accurate second order statistics. In practice, postulated statistics which are mismatched with respect to the real ones, are available. A mismatched covariance matrix model is introduced to facilitate a comparison for the recovery performance of the LMMSE estimation in different mismatch regimes. In this context, a performance comparison between SVT and LMMSE estimation for different levels of mismatch is presented across different noise regimes.

### 3.5.1 Comparison with the performance limits

This section presents a comparison between the SVT algorithm, the LMMSE estimation and the OPTA for a signal to noise ratio value of SNR= 20 dB. Figure 3.6 shows the performance, measured by NMSE, for the SVT-based recovery compared to the performance of the LMMSE estimator with access to perfect second order statistics, and to the theoretical limit given by the OPTA. As expected, the LMMSE estimation outperforms the SVT algorithm across all sampling regimes. However, the comparison with the OPTA highlights the sampling regimes in which the difference in recovery performance is larger. Precisely, for small values of $\gamma$, the difference between the SVT recovery and the OPTA is larger than in the case when $\gamma$ is large. Consequently, the performance of SVT is closer to the optimal recovery when the

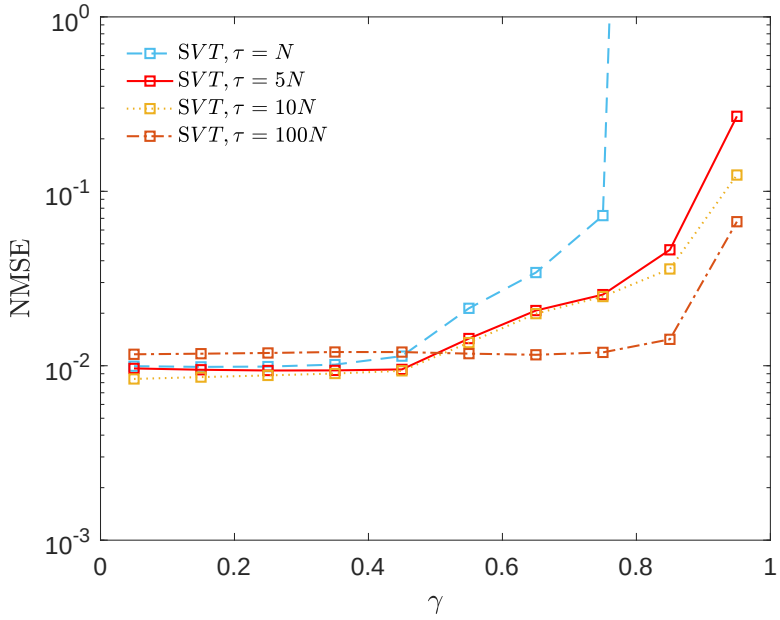



Figure 3.7. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT with different values of $\tau$, when SNR = 20 dB.

number of missing entries is large. In contrast, the difference between the LMMSE estimation and the SVT algorithm is relatively constant across all sampling regimes. Note that in comparison with SVT, the LMMSE estimation requires prior knowledge, i.e., the covariance matrix. In Figure 3.6, the numerical results for the LMMSE estimation assume that perfect second order statistics are available. In practice, this assumption is difficult to satisfy. For the SVT algorithm, it is noted in [19] that the value of the threshold $\tau$ significantly impacts the recovery performance across different sampling regimes. In view of this, the following section presents a performance comparison for the SVT recovery with different values of $\tau$.

### 3.5.2 Impact of $\tau$ on the recovery performance

A performance comparison for the SVT recovery with different values of the threshold is presented in this section. As presented in [38] and [19], the value of $\tau = 5N$ is not optimal. In [38] the value of the threshold is tuned via numerical optimization. Moreover, numerical results presented in [19] show that the optimal value for $\tau$ depends on the sampling regime. Large values of $\tau$ yield a good performance when a small number of observations is available. Conversely, smaller values of $\tau$ yield a good performance when a large number of observations is available.

Figure 3.7 shows the performance of the SVT recovery for different values of $\tau$ when SNR = 20 dB. The value of $\tau$ impacts the recovery performance across all sampling regimes and induces a performance trade-off. For instance, the value

$\tau = 5N$ suggested in [39] exhibits a compromise between the recovery performance on the high sampling regime and the low sampling regime. By changing the value of $\tau$, the recovery can be improved on a particular regime but at the expense of decreasing the performance on another sampling regime. This is discussed in [19] where it is noted that optimizing $\tau$ for a particular dataset is hard in general. Consequently, it is difficult to obtain the optimal SVT recovery in a practical setting. However, the same limitation applies to the LMMSE estimation because the optimal recovery is guaranteed only when access to perfect second order statistics is available. When the available statistics are mismatched with respect to the actual ones, the performance of the LMMSE recovery is suboptimal.

### 3.5.3 Impact of mismatched statistics on the recovery performance

This section presents a numerical comparison of the recovery performance of the LMMSE estimator when the perfect second order statistics are not available. The case in which the LMMSE estimator is optimal is also included for benchmarking purposes.

Since the performance of the LMMSE estimator depends on the covariance matrix $\mathbf{\Sigma}$, a mismatched covariance matrix model [104] and [105] is introduced to account for the difference between the postulated and actual statistics. Specifically, the postulated covariance matrix is given by

$$\mathbf{\Sigma}^* = \mathbf{\Sigma} + \frac{1}{\text{SMR}} \frac{\|\mathbf{\Sigma}\|_F^2}{\|\mathbf{\Delta}\|_F^2} \mathbf{\Delta}, \tag{3.39}$$

where $\mathbf{\Sigma}$ is the actual covariance matrix in (3.3), $\mathbf{\Delta} = \mathbf{H}\mathbf{H}^T$ with $\mathbf{H} \in \mathbb{R}^{M \times M}$ any matrix whose entries are distributed as $\mathcal{N}(0, 1)$. The strength of the mismatch is determined by the signal to mismatch ratio (SMR). Note that the SMR is defined such that for SMR $= 1$ the norm of the mismatch is equal to the norm of the real covariance matrix

$$\|\mathbf{\Sigma}\|_F^2 = \left\| \frac{1}{\text{SMR}} \frac{\|\mathbf{\Sigma}\|_F^2}{\|\mathbf{\Delta}\|_F^2} \mathbf{\Delta} \right\|_F^2. \tag{3.40}$$

As depicted in Figure 3.8, the LMMSE estimator gives better recovery performance when SMR $\geq$ 100. However, when SMR=10 and $\gamma \leq 0.55$ the SVT algorithm outperforms the LMMSE estimator. Moreover, the SVT provides a better recovery for SMR=1 for almost all values of $\gamma$. In view of this, the LMMSE estimation requires accurate second order statistics which is an unrealistic assumption in a practical scenario. Numerical results in Figure 3.8 assume a signal to noise ratio of 20 dB.

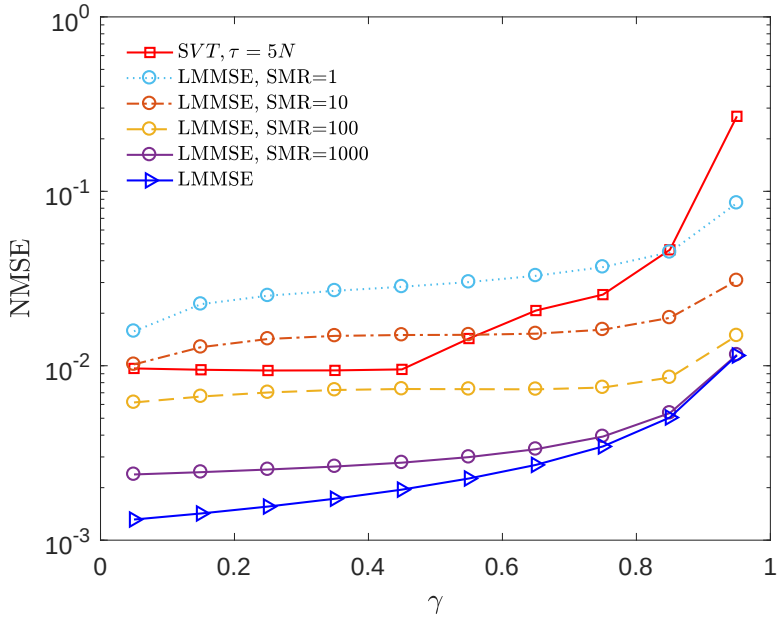



Figure 3.8. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and LMMSE estimation, for different levels of mismatch, when SNR = 20 dB.

The following section presents a recovery performance comparison for different SNR regimes.

### 3.5.4 Performance comparison for different SNR regimes

Numerical results presented up to this point assume a high SNR regime, i.e., SNR = 20 dB. In the following, a comparison between the SVT algorithm and the LMMSE estimation with and without access to perfect second order statistics is illustrated for other SNR regimes, i.e., the low SNR regime (0 dB), the medium SNR regime (10 dB) and the approximately noiseless case in which SNR = 50 dB.

Figure 3.9 depicts the recovery performance comparison for the SVT algorithm and the LMMSE estimation for different levels of mismatch, in the low SNR regime. The performance of the optimal LMMSE estimation is also included. In comparison to the high SNR regime i.e., SNR = 20 dB, the difference in performance between the optimal LMMSE estimator and SVT is larger. Moreover, the impact of mismatched statistics is not as significant as before. For a value of SMR = 100, the recovery performance is close to the optimal case. It is also interesting to notice when the number of missing entries is small, the impact of the mismatched statistics is more significant than in the case when $\gamma$ is large. This observation holds for both SNR values with a even more significant impact in the high SNR regime.

The performance comparison between SVT and LMMSE estimation in the medium SNR regime is depicted in Figure 3.10. Similarly to the low SNR regime, the LMMSE

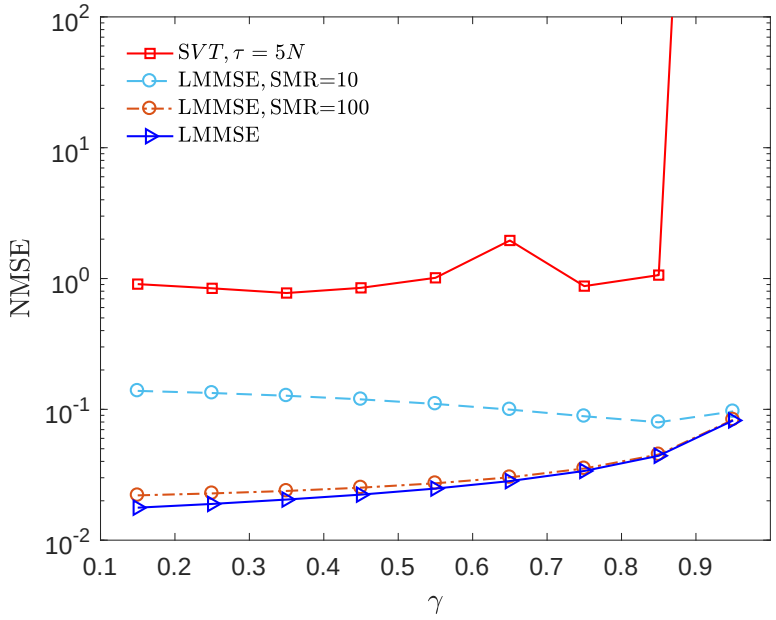



Figure 3.9. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and LMMSE estimation for different levels of mismatch, when SNR = 0 dB.

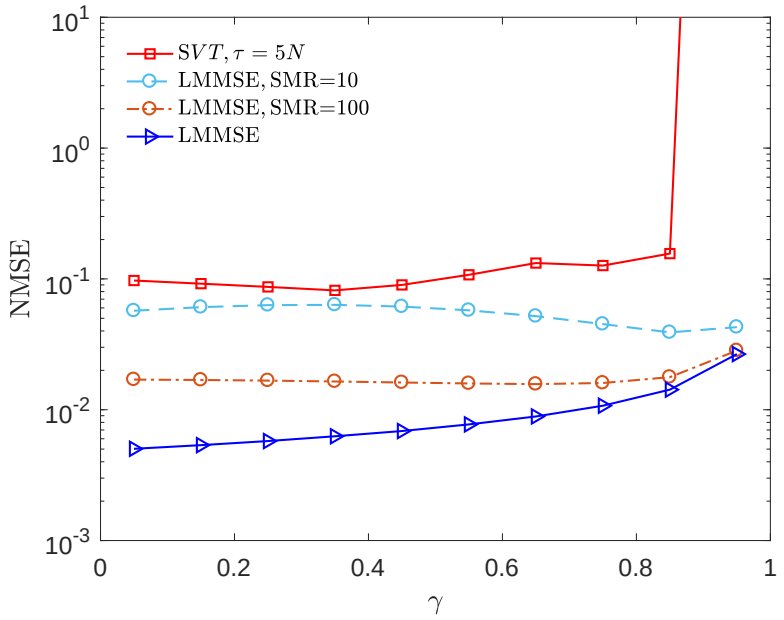



Figure 3.10. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and LMMSE estimation for different levels of mismatch, when SNR = 10 dB.

estimation outperforms the SVT recovery for $\tau = 5N$ even when mismatched statistics are available. However, the difference in performance is smaller in the medium SNR regime compared to the low SNR regime. In fact, the performance of

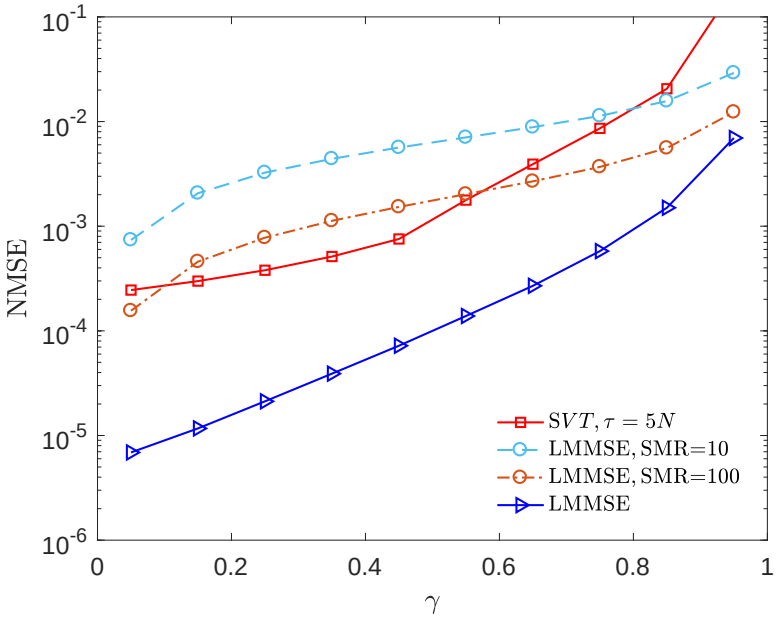



Figure 3.11. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and LMMSE estimation for different levels of mismatch, when SNR = 50 dB.

the SVT algorithm is closer to the optimal LMMSE estimation when the noise level is lower. This is expected because the SVT recovery is designed for the noiseless case [39]. Unfortunately, the recovery in the noiseless case is not possible using the LMMSE estimation because the covariance matrix is singular. In order overcome this limitation, a value of SNR = 50 dB is analyzed. The recovery performance of SVT and LMMSE estimation in the approximately noiseless case is depicted in Figure 3.11. Different levels of mismatch are considered for the LMMSE estimation.

Interestingly, the SVT based recovery outperforms the LMMSE estimation for a SMR value of 10 in most of the sampling regimes. Moreover, for $0.1 \leq \gamma \leq 0.55$ the matrix completion based recovery performs better than LMMSE with SMR=100. As expected, the LMMSE estimator with access to perfect second order statistics is the optimal recovery method in all sampling regimes.

Comparing the impact of the mismatched statistics across different SNR regimes it is interesting to notice that the lower the level of noise the more important is the precision of the postulated statistics. The difference between the optimal LMMSE and the mismatched cases increases with the value of SNR. In the same time, the precision of the SVT-based recovery increases with the value of SNR. Therefore, the matrix completion based recovery is an alternative to LMMSE estimation on high SNR regimes when the postulated statistics are not accurate. In a practical setting, the level of mismatch between the postulated statistics and the actual ones is hard to determine. As a consequence, the precision of the LMMSE estimate based on the

postulated statistics is unknown. In contrast, the performance of the SVT recovery, for a value of $\tau = 5N$, depends only on the SNR regime. In a realistic scenario, the value of the SNR depends on the quality of the sensors. High precision sensors and good communication infrastructure determines a high SNR value. In contrast, low quality sensors and a noisy communication channel result in a low SNR regime.

## 3.6 Conclusions

This chapter presented a numerical simulation study in which the performance of the SVT algorithm is compared with the LMMSE estimation under different SNR regimes and different levels of mismatch in the postulated statistics. Moreover, the optimal performance theoretically attainable by any estimator is defined using the Rate-Distortion theory for a multivariate Gaussian source. The numerical comparison with the OPTA depicted in Figure 3.6 shows that the performance recovery of both methods is largely suboptimal when the number of available entries is large and it is closer to the theoretical limit when most of the entries are missing.

For the SVT algorithm, the recommended value of the threshold $\tau$ represents a compromise between different sampling regimes. One the other side, when mismatch is considered for the second order statistics available, the performance of the LMMSE estimation decreases significantly. As an example, in the almost noiseless case the SVT aglorithm outperforms the LMMSE with SMR=100 for $0.15 \leq \gamma \leq 0.55$. The impact of mismatched statistics for the performance of the LMMSE estimator is more significant when the noise level is low. In contrast, for values of SNR$<$ 20 dB the LMMSE based recovery outperforms the SVT algorithm even for SMR$=$ 10. In view of this, the SVT recovery is a better alternative for recovering missing data when the noise level is low, while the LMMSE estimator performs better for high levels of noise in the available observations.

Note that while the numerical results presented in this chapter use real data from a distribution system, the applicability of both estimation methods is not limited to this type of data. In fact, the applicability of the SVT algorithm extends to any dataset that generates a low rank or approximately low rank matrix. Furthermore, the LMMSE estimator is the best linear unbiased estimator and it gives optimal recovery when the data follows a Gaussian distribution.

The main limitation of the SVT recovery is the lack of guidelines for tuning the value of $\tau$ depending on the noise level and the sampling regime. On the other hand, the performance of the LMMSE estimator varies significantly with the quality of the prior knowledge. Therefore, the LMMSE estimator is not robust to imperfect prior knowledge and has limited applicability in a practical setting for the recovery of missing data.

In this context, the following chapter introduces a novel algorithm for recovering missing data. The proposed approach is a modified version of the SVT algorithm that optimizes the threshold $\tau$ at each iteration based on the prior knowledge incorporated in the structure of the algorithm.

# Chapter 4

# A novel algorithm for robust recovery of missing data

## 4.1 Introduction

This chapter introduces a new algorithm for recovering missing data in high-dimensional datasets. The proposed algorithm is referred to as Bayesian SVT (BSVT). Numerical results presented in Section 3.5 emphasize the limitations imposed by the two state-of-the-art missing data recovery algorithms studied. Specifically, for the SVT algorithm the optimization of the threshold $\tau$ is difficult and therefore the recovery performance is suboptimal across different sampling regimes. In practice, the value of $\tau$ is obtained by trial and error [38]. In the case of the LMMSE estimator, the recovery performance relies on accurate prior knowledge which is not available in a practical setting.

The proposed algorithm addresses the limitations of both approaches (SVT and LMMSE). In particular, the new recovery method is a modified version of the SVT algorithm which incorporates an optimization step for the threshold $\tau$ at each iteration that exploits the same prior knowledge as the LMMSE estimator.

The performance of the BSVT algorithm is tested through numerical simulations. Both the convergence speed and the recovery performance are assessed in different noise and sampling regimes. Moreover, practical settings in which the prior knowledge is not accurate and the sampling pattern is non-uniform are also evaluated. In view of this, the proposed recovery method provides robust performance in realistic scenarios.

In the following, the system model is identical to the one used in Chapter 3, which is described in Section 3.2. The new recovery algorithm is introduced in Section 4.2 and theoretical guarantees for convergence are provided in Section 4.3. A numerical comparison between SVT and the novel approach is presented in Section 4.4. To that end, a non-uniform sampling model is introduced in Section 4.4 to account

for structure in the observation pattern. Finally, the conclusions are presented in Section 4.5.

## 4.2 The Bayesian Singular Value Thresholding Algorithm

This section introduces a novel algorithm for recovering missing data that incorporates prior knowledge in the form of the second order statistics. Precisely, the new approach is a modified version of the SVT algorithm that optimizes the threshold $\tau$ at each iteration based on additional prior knowledge. The numerical simulation results presented in [19] and in Section 3.5.2 show that the recommended threshold for the SVT algorithm, i.e., $\tau = 5N$, performs poorly when the number of missing entries is large. The choice of $\tau$ governs the performance trade-off between the high and low sampling regimes. Large values of $\tau$ yield a good performance when a small number of observations is available. Conversely, smaller values of $\tau$ yield a good performance when a large number of observations is available. Unfortunately, finding the optimal threshold for the SVT algorithm is still an open problem because the input matrix for the soft-thresholding step is sparse. In general, the value of the threshold for soft-thresholding based recovery algorithms is obtained via numerical optimization in [19] and [38].

### 4.2.1 Soft-thresholding parameter

The main shortcoming of the SVT algorithm is the lack of guidelines for tuning the threshold $\tau$. Numerical results in Section 3.5.2 show that the $5N$ value proposed in [39] is not optimal for every scenario. In order to provide better recovery it is essential to tune the value of $\tau$ for each iteration of the algorithm. In SVT the soft-thresholding operator is applied on a sparse matrix for which there are no guidelines for the tuning process. However, the same soft-thresholding operator, $D_\tau$, is used in a different framework for denoising in [38], [106], and [107]. The main difference between the two frameworks is that for denoising the soft-thresholding operator is applied on a complete matrix.

In general, measuring the performance of the soft-thresholding operator for different values of $\tau$, in the MSE sense, requires access to the true matrix $\mathbf{M}$. However, for the denoising framework, an unbiased estimator for the MSE, namely Stein's unbiased risk estimate (SURE) is presented in [108]. Moreover, a closed-form expression for SURE is presented in [109] for the soft-thresholding operator which facilitates the optimization of $\tau$ in a denoising framework. In order to calculate

SURE in the recovery setting, the missing entries are initially estimated using the prior knowledge incorporated in the structure of the new algorithm.

### 4.2.2 Exploiting second order statistics

In order to overcome the limitation imposed by the sparse structure of the matrix $\mathbf{Y}^{(k)}$, the proposed algorithm estimates the missing entries prior to the soft-thresholding step using the second order statistics incorporated as prior knowledge. Thus, the available prior knowledge is exploited to produce an estimate of the entries not contained in $\Omega$. In this case, at each iteration $k$ of the proposed algorithm the matrix $\mathbf{Z}^{(k)}$ is computed as

$$\mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} + \mathbf{L}^{(k)}, \tag{4.1}$$

where $\mathbf{Y}^{(k)}$ is defined as in the SVT algorithm and $\mathbf{L}^{(k)}$ is the LMMSE estimate given by

$$\mathbf{L}^{(k)} = P_{\Omega^c}(\boldsymbol{\mu}) + \boldsymbol{\Sigma}^*_{\Omega^c\Omega}\boldsymbol{\Sigma}^{*-1}_{\Omega\Omega}(P_{\Omega}(\mathbf{Y}^{(k)}) - P_{\Omega}(\boldsymbol{\mu})), \tag{4.2}$$

where $\Omega$ is the set of observed entries, $\Omega^c$ is the set of missing entries, $\boldsymbol{\Sigma}^*_{\Omega^c\Omega}$ is the covariance matrix between the entries in $\Omega^c$ and the entries in $\Omega$ and $\boldsymbol{\Sigma}^*_{\Omega\Omega}$ is the covariance matrix of the entries in $\Omega$. Therefore, the unknown entries are estimated using the LMMSE-based recovery at each iteration $k$. The result is a complete matrix $\mathbf{Z}^{(k)}$ for which the optimization of the threshold $\tau$ is feasible.

### 4.2.3 Optimization of thresholding parameter

The closed-form expression for the risk estimator provided in [109] is incorporated into the proposed algorithm to facilitate the optimization of the thresholding parameter. The use of SURE is made possible by the addition of the LMMSE step, which provides a linear estimate for the entries in $\Omega_c$. This ensures that the matrix provided as input to the soft-thresholding step is complete and the optimization of $\tau$ is solvable as a denoising problem. In this context, the performance of the soft-thresholding operator can be estimated when the input matrix $\mathbf{Z}$ accepts the following model [109]

$$\mathbf{Z} = \mathbf{M} + \mathbf{W}, \tag{4.3}$$

where the entries of $\mathbf{W}$ are

$$(\mathbf{W})_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\mathbf{Z}}), \tag{4.4}$$

where $\sigma^2_{\mathbf{Z}}$ is the variance of the $(\mathbf{W})_{i,j}$ entries with $i \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, N\}$. In this setting, the SURE [108] is given by

$$\begin{aligned} \mathrm{SURE}(D_\tau)(\mathbf{Z}) = &- MN\sigma^2_{\mathbf{Z}} + \sum_{i=1}^{\min\{M,N\}} \min\{\tau^2, \sigma_i^2(\mathbf{Z})\} \\ &+ 2\sigma^2_{\mathbf{Z}}\mathrm{div}(D_\tau(\mathbf{Z})), \end{aligned} \tag{4.5}$$

where $\sigma_i(\mathbf{Z})$ is the $i$-th singular value of $\mathbf{Z}$ for $i \in \{1, 2, \ldots, N\}$. A closed-form expression for the divergence of this estimator is obtained in [109]. For the case in which $\mathbf{Z} \in \mathbb{R}^{M\times N}$ the divergence is given by

$$\begin{aligned} \mathrm{div}(D_\tau(\mathbf{Z})) = &\sum_{i=1}^{\min\{M,N\}} \left[\mathbb{1}(\sigma_i(\mathbf{Z}) > \tau) + |M - N|\frac{(\sigma_i(\mathbf{Z}) - \tau)_+}{\sigma_i(\mathbf{Z})}\right] \\ &+ 2\sum_{i\neq j, i,j=1}^{\min\{M,N\}} \frac{\sigma_i(\mathbf{Z})(\sigma_i(\mathbf{Z}) - \tau)_+}{\sigma_i^2(\mathbf{Z}) - \sigma_j^2(\mathbf{Z})}, \end{aligned} \tag{4.6}$$

when $\mathbf{Z}$ has no repeated singular values and is zero otherwise, where $\mathbb{1}(\cdot)$ denotes the indicator function defined as

$$\mathbb{1}(\sigma_i(\mathbf{Z}) > \tau) = \begin{cases} 1, & \text{if } \sigma_i(\mathbf{Z}) > \tau, \\ 0, & \text{if } \sigma_i(\mathbf{Z}) \leq \tau. \end{cases} \tag{4.7}$$

Therefore, combining (4.5) and (4.6) gives a closed-form expression for the performance of the soft-thresholding operator for different values of $\tau$ and $\sigma^2_{\mathbf{Z}}$.

The proposed algorithm approximates $\sigma^2_{\mathbf{Z}}$ with the weighted sum of the noise in $\Omega$ and in $\Omega^c$, i.e., $\sigma^2_{\mathbf{Z}^{(k)}}$ is calculated as

$$\sigma^2_{\mathbf{Z}^{(k)}} = \frac{\|\mathbf{Y}^{(k)} - P_\Omega(\mathbf{M})\|_F^2 + |\Omega^c| D_{\mathrm{LMMSE}}}{MN}, \tag{4.8}$$

where $D_{\mathrm{LMMSE}}$ represents the average noise per entry in $\Omega^c$ and is given by

$$D_{\mathrm{LMMSE}} = \frac{\mathrm{tr}(\mathbf{\Sigma}^*_{\Omega^c\Omega^c} - (\mathbf{\Sigma}^*_{\Omega^c\Omega}\mathbf{\Sigma}^{*-1}_{\Omega\Omega}\mathbf{\Sigma}^*_{\Omega\Omega^c}))}{|\Omega^c|}. \tag{4.9}$$

The optimal threshold for the matrix $\mathbf{Z}^{(k)}$ is denoted by $\tau^{(k)}$ and it is calculated using

$$\tau^{(k)} = \arg\min_\tau \mathrm{SURE}(D_\tau)(\mathbf{Z}^{(k)}), \tag{4.10}$$

where $\sigma^2_{\mathbf{Z}^{(k)}}$ is given by (4.8).

Note that the cost function in (4.10) is quasiconvex and is solved using standard optimization tools [109] over a predefined interval $[\tau_{\min}, \tau_{\max}]$.

Therefore, the iterations of the proposed algorithm are

$$\begin{cases} \mathbf{X}^{(k)} = D_{\tau^{(k-1)}}(\mathbf{Z}^{(k-1)}), \\ \mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_b\Big(P_\Omega(\mathbf{M}) - P_\Omega(\mathbf{X}^{(k)})\Big), \\ \mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} + \mathbf{L}^{(k)}, \\ \tau^{(k)} = \arg\min_{\tau} \mathrm{SURE}(D_\tau)(\mathbf{Z}^{(k)}), \end{cases} \tag{4.11}$$

where the $D_\tau$ is defined by (3.14) and the step size $\delta_b$ is similar to the step size $\delta_s$ in the SVT algorithm. The initial conditions are $\mathbf{Z}^{(0)} = \mathbf{0}$, $\mathbf{Y}^{(0)} = \mathbf{0}$ and $\tau^{(0)} = 0$. The stopping criteria is the same as that of the SVT algorithm, namely

$$\frac{\|P_\Omega(\mathbf{X}^{(k)}) - P_\Omega(\mathbf{M})\|_F}{\|P_\Omega(\mathbf{M})\|_F} \leq \epsilon. \tag{4.12}$$

A more detailed description of the proposed algorithm is presented in Algorithm 1.

**Algorithm 1** Bayesian Singular Value Thresholding

**Require:** set of observations $\Omega$, observed entries $P_\Omega(\mathbf{R})$, mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}^*$, step size $\delta_b$, tolerance $\epsilon$ and maximum iteration count $k_{\max}$

**Ensure:** $\widehat{\mathbf{M}}_{\mathrm{BSVT}}$

Set $\mathbf{Y}^0 = \mathbf{0}$
Set $\mathbf{Z}^0 = \mathbf{0}$
Set $\tau = 0$
Set $\Omega^c = \{1, 2, ..., M\} \times \{1, 2, ..., N\} \setminus \Omega$
**for** $k = 1$ to $k_{\max}$ **do**
    Compute $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{svd}(\mathbf{Z}^{(k-1)})$
    Set $\mathbf{X}^{(k)} = \sum_{j=1}^{\min\{M,N\}} \max\{0, \sigma_j(\mathbf{Z}^{(k-1)}) - \tau^{(k-1)}\}\mathbf{u}_j\mathbf{v}_j$
    **if** $\|P_\Omega(\mathbf{X}^{(k)}) - P_\Omega(\mathbf{R})\|_F / \|P_\Omega(\mathbf{R})\|_F \leq \epsilon$ **then break**
    **end if**
    Set $\mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_b\Big(P_\Omega(\mathbf{R}) - P_\Omega(\mathbf{X}^{(k)})\Big)$
    Set $\mathbf{L}^{(k)} = P_{\Omega^c}(\boldsymbol{\mu}) + \boldsymbol{\Sigma}^*_{\Omega^c\Omega}\boldsymbol{\Sigma}^{*-1}_{\Omega\Omega}(\mathbf{Y}^{(k)} - P_\Omega(\boldsymbol{\mu}))$
    Set $\mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} + \mathbf{L}^{(k)}$
    Set $\sigma^2_{\mathbf{Z}^{(k)}} = (\|\mathbf{Y}^{(k)} - P_\Omega(\mathbf{R})\|^2_F + |\Omega^c| D_{\mathrm{LMMSE}})/MN$
    Set $\tau^{(k)} = \arg\min_{\tau} \mathrm{SURE}(D_\tau)(\mathbf{Z}^{(k)})$
**end for**
Set $\widehat{\mathbf{M}}_{\mathrm{BSVT}} = \mathbf{X}^{(k)}$

The main advantage of the new algorithm is that the threshold is optimized at each iteration facilitated by the prior knowledge incorporated into the structure of the algorithm, i.e., the covariance matrix. First, an initial guess of the unavailable entries is formed at each iteration $k$ based on $\mathbf{Y}^{(k)}$ and the available covariance matrix $\boldsymbol{\Sigma}^*$. The results are aggregated in the matrix $\mathbf{Z}^{(k)}$ which is approximated by the model in (4.3). In this case, an estimate of the noise level, $\sigma^2_{\mathbf{Z}^{(k)}}$, is needed

to compute SURE. The optimal value of $\tau^{(k)}$ for $\mathbf{Z}^{(k)}$ is obtained by minimizing $\text{SURE}(D_\tau)(\mathbf{Z}^{(k)})$ in (4.5). Admittedly, the optimization of the threshold requires a priori knowledge of the second order statistics. Therefore, the new approach requires additional knowledge that is not necessary when using the SVT algorithm. That being said, the SVT algorithm requires setting the value for the threshold which in general is difficult to tune. The same amount of prior knowledge, i.e., covariance matrix, is required by the LMMSE estimator. Numerical results presented in Section 3.5.4 show that when the postulated statistics are not accurate, the estimation error (NMSE) for the LMMSE-based recovery increases by an order of magnitude. For the proposed algorithm, the trade-off between the performance and the accuracy of the prior knowledge is studied in Section 4.4.3. Moreover, the following section provides theoretical guarantees for the proposed algorithm to outperform the SVT recovery.

## 4.3 Theoretical guarantees for convergence of the BSVT algorithm

This section presents theoretical guarantees for the convergence of the BSVT algorithm. Moreover, the theoretical analysis in this section compares the performance of the BSVT and SVT algorithms, presented in Section 4.2 and Section 3.3 respectively, and provides conditions for which the proposed algorithm theoretically outperforms the SVT recovery at each iteration.

**Theorem 1.** *Let $\mathbf{X}_b^{(k)}$ be the estimate produced by the BSVT algorithm at iteration $k$ and $\mathbf{X}_s^{(k)}$ be the estimate produced by the SVT algorithm at iteration $k$ with $\mathbf{L}^{(k)}$ denoting the LMMSE estimation output at iteration $k$ as defined in (4.2). If the following holds:*

$$tr\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \geq tr\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big), \tag{4.13}$$

*and*

$$\lim_{k\to\infty} \sigma_{\min}(\mathbf{L}^{(k)}) = 0, \tag{4.14}$$

*then*

$$\|\mathbf{X}_b^{(k)} - \mathbf{M}\|_F^2 \leq \|\mathbf{X}_s^{(k)} - \mathbf{M}\|_F^2. \tag{4.15}$$

*Proof.* See Appendix A. □

**Corollary 1.1.** *Let $k_s$ be the last iteration of the SVT algorithm and denote by $\mathbf{X}_s^{(k_s)}$ the output of the SVT algorithm. For the case in which $\mathbf{X}_b^{(k)}\mathbf{M}^T$ and $\mathbf{X}_s^{(k)}\mathbf{M}^T$ are Hermitian, the condition (4.13) is equivalent to*

$$0 \leq \sigma_{\max}(\mathbf{L}^{k_s})(\|\mathbf{X}_s^{k_s}\|_* + \varepsilon) + \sum_{i=1}^{\min\{M,N\}} \sigma_i(\mathbf{X}_s^{k_s})(\sigma_{\min\{M,N\}-i+1}(\mathbf{X}_s^{k_s}) - \sigma_i(\mathbf{X}_s^{k_s}) + \varepsilon), \tag{4.16}$$

*where $\varepsilon$ is a small numerical constant, $\sigma_i(\mathbf{X}_s^{(k_s)})$ denotes the ith singular value of the matrix $\mathbf{X}_s^{(k_s)}$ and $\sigma_{\max}(\mathbf{L}^{(k_s)})$ denotes the maximum singular value of the matrix $\mathbf{L}^{(k_s)}$.*

*Proof.* See Appendix A. □

Note that in Corollary 1.1 when the rank of the matrix $\mathbf{X}_s^{(k_s)}$, denoted by $r_s$, satisfies $r_s < \frac{\min\{M,N\}}{2}$, $\sigma_{\min\{M,N\}-i+1}(\mathbf{X}_s) = 0$ for $i \in \{1, 2, ..., \min\{M, N\}\}$ and (4.16) becomes

$$0 \leq \sum_{i=1}^{r_s} \sigma_i(\mathbf{X}_s^{k_s})\Big(\sigma_{\max}(\mathbf{L}^{k_s}) - \sigma_i(\mathbf{X}_s^{k_s})\Big) + \Big(\|\mathbf{X}_s^{k_s}\|_* + \sigma_{\max}(\mathbf{L}^{k_s})\Big)\varepsilon, \tag{4.17}$$

where $\varepsilon$ is a small numerical constant.

Theorem 1 provides a condition for the optimality of the BSVT recovery compared to that of the SVT algorithm. Another consequence of Theorem 1 is that when the inequality in (4.13) is satisfied and the SVT algorithm converges, the proposed algorithm also converges and more importantly, it provides a better approximation. In addition, when the recovery error of the BSVT algorithm lower bounds the recovery error of the SVT approach, the proposed algorithm converges under milder conditions than those of SVT. This phenomenon is observable in all the numerical results presented in the following section for the cases in which the proportion of missing entries is larger than 0.9.

When the matrices $\mathbf{X}_b^{(k)}\mathbf{M}^T$ and $\mathbf{X}_s^{(k)}\mathbf{M}^T$ are Hermitian, the inequality in (4.13) is equivalent to the inequality in (4.16). This inequality does not depend on the matrix $\mathbf{M}$, and therefore the condition is decoupled from particular data realizations.

The following section provides a numerical comparison between the BSVT and the SVT algorithms in terms of recovery error, number of iterations and computation time required. In addition, practical scenarios are addressed for the cases in which the prior knowledge is not accurate and the locations of the missing entries are not uniformly distributed.

## 4.4 Numerical Analysis

This section evaluates the performance of the BSVT algorithm using the LV dataset presented in Section 3.2 consisting of the $500 \times 500$ data matrix $\mathbf{M}$, containing the voltage measurements from the electricity distribution system. In this context, the data matrix is sampled uniformly at random in order to decide which set of entries is available and which is estimated. The sampling probability is given by

$$\Pr[(i, j) \in \Omega] = \frac{1}{MN}\mathbb{E}[|\Omega|]. \tag{4.18}$$

The sampling probability for each entry of the matrix is a function of the expected number of available entries. Let $\gamma$ be the expected value of the proportion of missing entries for the matrix $\mathbf{M}$, that is:

$$\gamma \triangleq 1 - \frac{1}{MN}|\Omega|. \tag{4.19}$$

The performance of the algorithms is measured in terms of NMSE given by

$$\text{NMSE} = \frac{\|\mathbf{M} - \widehat{\mathbf{M}}\|_F^2}{\|\mathbf{M}\|_F^2}, \tag{4.20}$$

where $\widehat{\mathbf{M}}$ is the output of the recovery algorithm. The performance of each recovery technique is averaged over one hundred realizations of $\Omega$ for each proportion of missing entries.

The performance of the proposed algorithm is evaluated numerically using four criteria: convergence speed, performance gain for the optimized threshold, robustness to imperfect prior knowledge and performance using non-uniform sampling.

The convergence speed is analyzed in Section 4.4.1 where a comparison in terms of number of iterations and computation time is presented for the BSVT algorithm and the SVT-based recovery.

The performance gain for the optimized threshold is assessed in Section 4.4.2 where the SVT-based recovery is compared with the BSVT algorithm when accurate second order statistics are available.

The robustness of the BSVT recovery to imperfect prior knowledge is also evaluated. A comparison between the SVT algorithm, the LMMSE estimator and the BSVT recovery is presented for different values of SMR in Section 4.4.3. The case in which perfect second-order statistics are available is also included.

The robustness of the BSVT recovery to different sampling patterns is evaluated using Markov-chain-based sampling. The numerical performance of the new algorithm is compared to the SVT algorithm in a practical scenario in which the positions of the missing entries are not uniformly distributed and the postulated statistics are not accurate in Section 4.4.4.

Moreover, all the aspects are tested across different SNR regimes: the almost noiseless case (SNR= 50 dB), the high SNR regime (SNR= 20 dB), the medium SNR regime (SNR= 10 dB) and the low SNR regime (SNR= 0 dB).

### 4.4.1 Numerical convergence comparison

In this section, the computational effort required by BSVT and SVT is compared. Based on the description of the algorithm, the computational cost per iteration increases in the case of the proposed algorithm when compared with that of SVT.

Table 4.1. Convergence performance comparison for the Uniform sampling case

| SNR | $\gamma$ | #Iters | | Time(s) | |
|---|---|---|---|---|---|
| | | BSVT | SVT | BSVT | SVT |
| 0 | 0.15 | 2 | 51 | 17.32 | 68.24 |
| | 0.85 | 4 | 30 | 32.06 | 53.62 |
| 10 | 0.15 | 4 | 191 | 52.38 | 111.59 |
| | 0.85 | 2 | 32 | 6 | 19.54 |
| 20 | 0.15 | 2 | 365 | 15.89 | 169.92 |
| | 0.85 | 2 | 92 | 13.8 | 84.17 |
| 50 | 0.15 | 18 | 485 | 271.47 | 100.42 |
| | 0.85 | 2 | 145 | 8.67 | 69.16 |

Table 4.2. Convergence performance comparison for the Markovian sampling case

| SNR | $\gamma$ | #Iters | | Time(s) | |
|---|---|---|---|---|---|
| | | BSVT | SVT | BSVT | SVT |
| 0 | 0.15 | 6 | 53 | 133.71 | 60.86 |
| | 0.85 | 2 | 20 | 17.1 | 22.89 |
| 10 | 0.15 | 13 | 175 | 293.74 | 104.08 |
| | 0.85 | 2 | 43 | 15.89 | 30.77 |
| 20 | 0.15 | 9 | 329 | 156.44 | 172.46 |
| | 0.85 | 2 | 80 | 16.59 | 12.6 |
| 50 | 0.15 | 9 | 446 | 152.96 | 72.72 |
| | 0.85 | 2 | 30 | 16.73 | 2.04 |

However, the additional LMMSE step and the computation of the optimal threshold $\tau$ significantly impact the speed of convergence.

Table 4.1 shows a convergence performance comparison between BSVT and SVT in terms of number of iterations and computation time for different values of SNR and different values of $\gamma$. The computational platform used is the Iceberg HPC cluster at The University of Sheffield. The results are averaged over ten realizations of $\Omega$. The column #Iters shows the minimum number of iterations required by each algorithm to achieve the corresponding NMSE values depicted in Figures 4.6, 4.7, 4.8 and 4.5. Time(s) denotes the time, measured in seconds, required to recover the missing

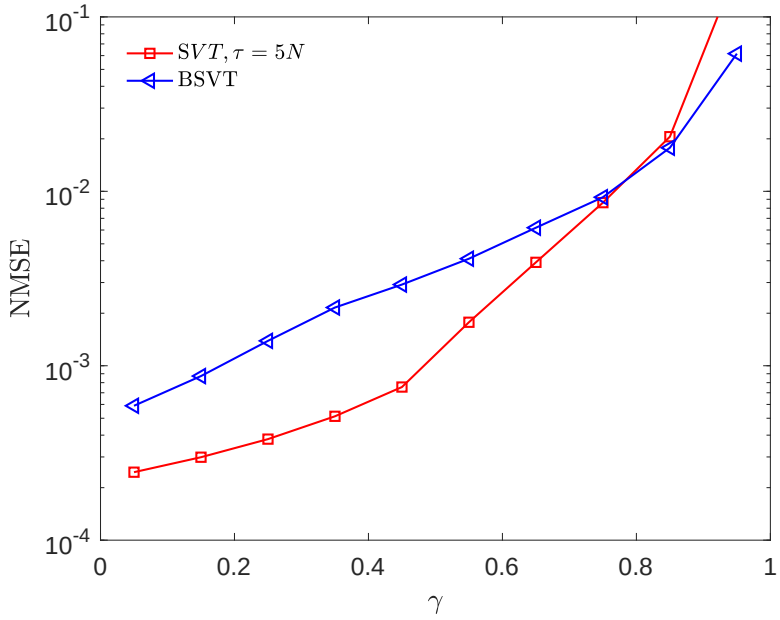



Figure 4.1. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and BSVT, when SNR = 50 dB.

entries in the matrix $\mathbf{M}$ for each case. A similar comparison for the Markovian sampling case is presented in Table 4.2. As expected, BSVT converges in fewer iterations but incurs in a higher computational cost per iteration when compared to SVT. Interestingly, the Markovian sampling case requires a larger number of iteration in most of the cases compared to the uniform sampling scenario. Moreover, BSVT requires fewer computation time in comparison to SVT in most of the cases with a large number of missing entries.

### 4.4.2 Performance of the optimized threshold

In this section, the performance of the new algorithm is compared to the SVT-based recovery using the same data matrix $\mathbf{M}$ and the same sets of available entries $\Omega$, for a particular proportion of missing entries $\gamma$ as defined in (4.19). The positions of the missing entries are sampled uniformly at random from the set of all entries.

Figure 4.1 shows the performance of the BSVT and SVT algorithms for recovering real LV data in the almost noiseless case. Not surprisingly, the SVT approach outperforms the proposed algorithm in this scenario because it is designed for the noiseless case. However, when the number of available entries is limited, i.e., $\gamma \geq 0.8$, the BSVT recovery performs better than the SVT algorithm. In addition, the problem of optimizing $\tau$ in the SVT recovery remains.

Figure 4.2 depicts the performance of both algorithms when applied in identical scenarios and SNR = 20 dB. Clearly, the optimized threshold and the Bayesian

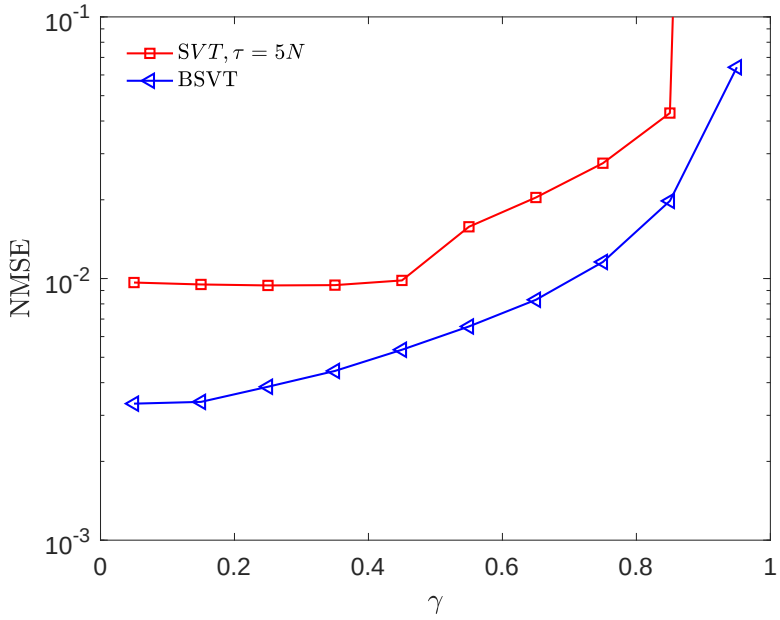



Figure 4.2. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and BSVT, when SNR = 20 dB.

estimation step increase the performance of the proposed algorithm in the high SNR regime when accurate second order statistics are available. When the postulated statistics, i.e., those available to the operator are identical to the real statistics, the BSVT algorithm provides a better performance for all values of $\gamma$. The gain in performance is larger when the proportion of missing entries is smaller than 0.4. Interestingly, the boost in performance is substantial in the region in which SVT is least efficient when compared to the fundamental limit that is presented in Section 3.5.1.

In Figure 4.3 the performance of the proposed algorithm is compared with the SVT recovery in the medium SNR regime. Interestingly, the BSVT approach outperforms the SVT algorithm by almost an order of magnitude in the NMSE. Also, the boost in performance is more significant when the proportion of missing entries is small and the SVT is least efficient compared to the theoretical limit.

The low SNR regime is depicted in Figure 4.4 and the new algorithm outperforms the SVT recovery by more than one order of magnitude in the NMSE. Note that the case in which $\gamma = 0.05$ is not included in Figure 4.4 because neither SVT nor BSVT converge to a low rank matrix in the experiments carried out.

In view of this, the BSVT recovery is suitable for the scenarios in which the observed entries are noisy. Moreover, the difference in recovery performance between SVT and BSVT increases as the level of noise increases. While in the almost noiseless case SVT performs better is most of the missing data regimes, as the value of SNR decreases the proposed algorithm outperforms the SVT recovery for all values of $\gamma$.

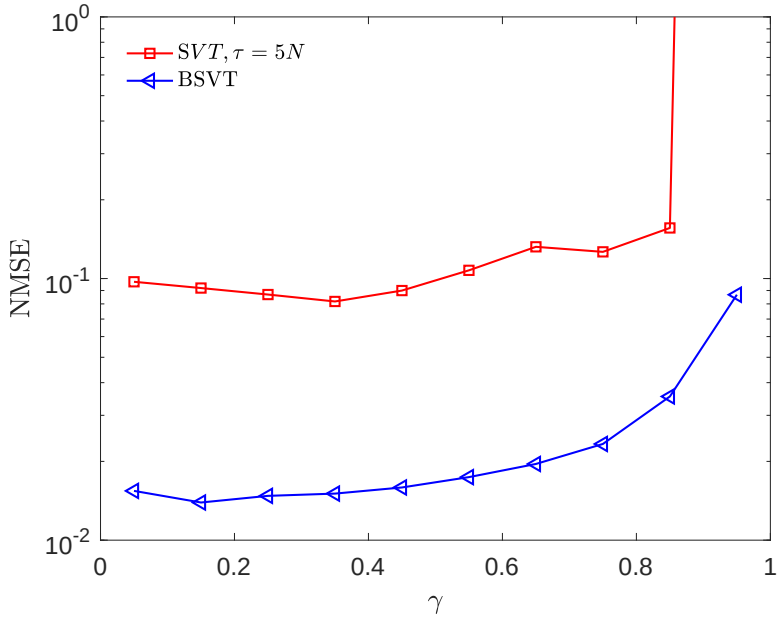



Figure 4.3. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and BSVT, when SNR = 10 dB.

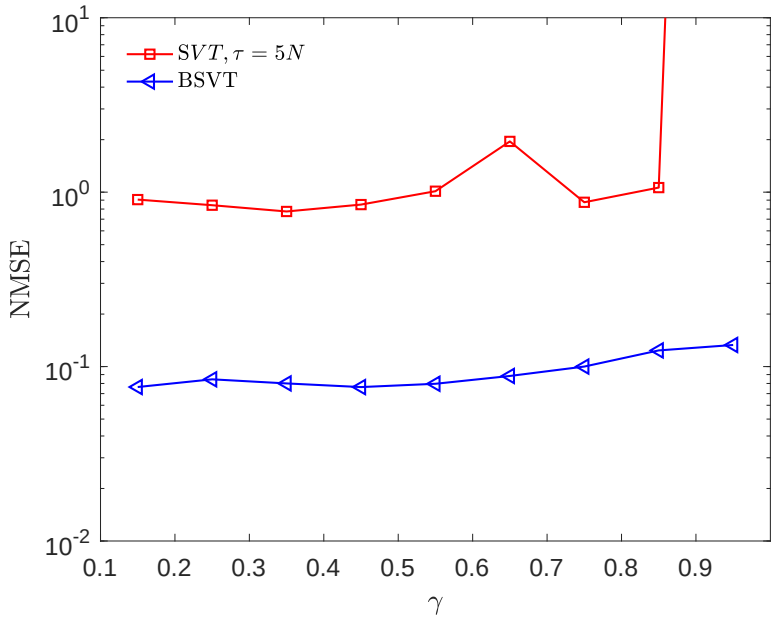



Figure 4.4. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT and BSVT, when SNR = 0 dB.

It is worth noting that the fast degradation with respect to observation noise of SVT reported in [39] is also corroborated by the simulations in this work.

However, numerical results in this section assume that the prior knowledge available is accurate. In practical scenarios the postulated and actual statistics are different. The impact of mismatched statistics is considered in the following section.

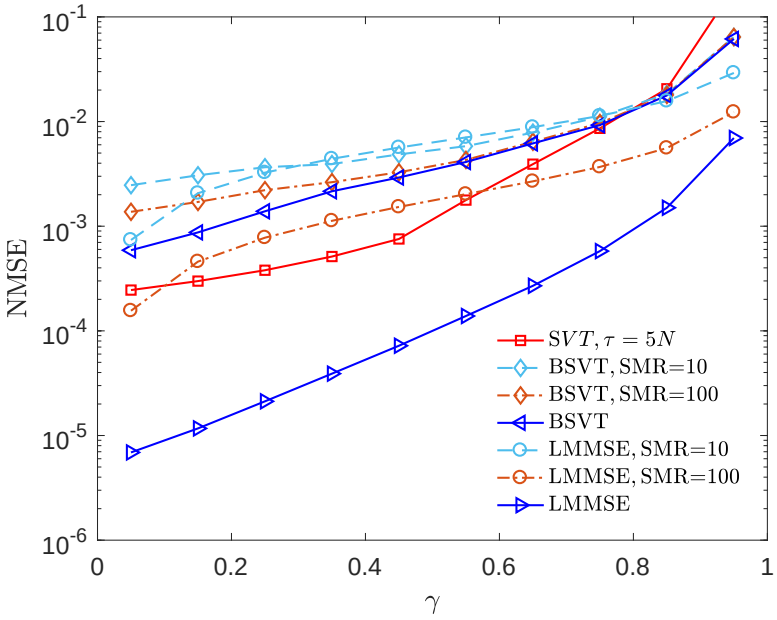



Figure 4.5. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT, LMMSE estimation and BSVT for different levels of mismatch, when SNR = 50 dB.

### 4.4.3 Performance using mismatched statistics

To address the problem of missing data recovery in a realistic scenario, a level of mismatch between the real covariance matrix and the one available to the operator is considered. The mismatch covariance matrix model presented in Section 3.5.3 is also used in this section to assess the sensibility of the proposed algorithm to inaccurate prior knowledge. Hence, the LMMSE estimator and the BSVT algorithm are compared in the no-mismatch regime and for a SMR value of 100 and 10. The performance of the SVT-based recovery is included as a benchmark for comparing rank minimization-based approaches.

A comparison between SVT, BSVT and the LMMSE estimator for different values of mismatch is presented in Figure 4.5 when SNR = 50 dB. In the almost noiseless case, SVT outperforms BSVT which highlights that SVT is well suited for noiseless problems. This is also noted in the numerical results presented in Figure 4.1. On the other hand, the mismatched statistics have a significant impact on the recovery error for the LMMSE approach. Specifically, for a value of SMR=100 the performance of the LMMSE estimator decreases by more than an order of magnitude in the NMSE compared with the no mismatch case. In contrast, the decrease in performance is less significant for the BSVT recovery for the same level of mismatch. In fact, when $\gamma \geq 0.8$ the performance of the new algorithm is not affected by the mismatch introduced in the prior knowledge and for $\gamma < 0.8$ the decrease in performance is less significant compared to the LMMSE-based recovery. In view of this, the proposed

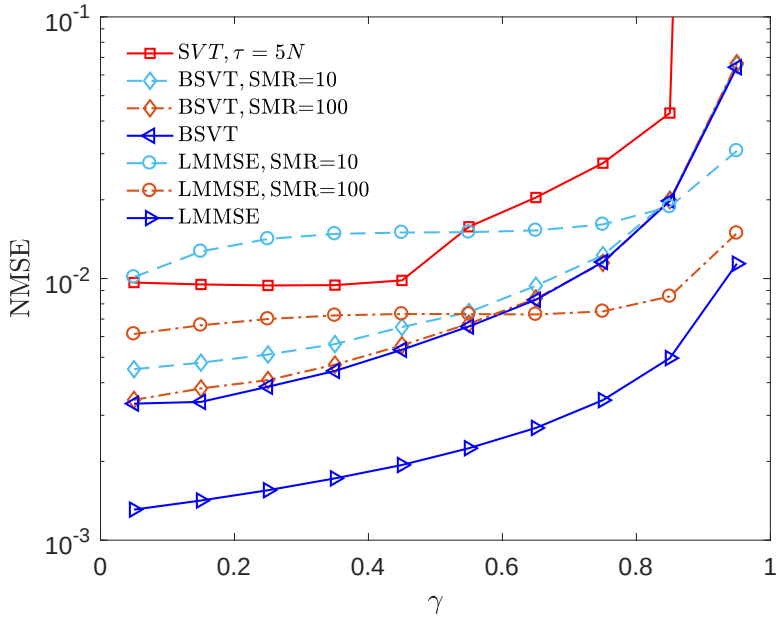



Figure 4.6. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT, LMMSE estimation and BSVT for different levels of mismatch, when SNR = 20 dB.

approach is more robust to mismatched statistics compared to the LMMSE estimator in the almost noiseless case.

Figure 4.6 depicts the performance of the SVT, BSVT, and the LMMSE estimator when mismatched second order statistics are available and SNR = 20 dB. In comparison with the almost noiseless case, in the high SNR regime, the mismatched statistics have less impact on the recovery error. However, the proposed algorithm is still robust to mismatch in the second order statistics while the performance of the LMMSE estimator decreases by an order of magnitude. In contrast, the performance of the BSVT algorithm does not change significantly when mismatch occurs. In fact, the BSVT algorithm gives better recovery than the SVT-based recovery in all mismatch regimes throughout the range of $\gamma$. In comparison with the LMMSE estimation, the BSVT algorithm performs better for SMR = 100 when $\gamma \leq 0.65$. Furthermore, for SMR = 10 the proposed approach is the best performing recovery method for almost all values of $\gamma$.

Figure 4.7 depicts the performance of the SVT, BSVT, and the LMMSE estimator for different values of mismatch, when SNR = 10 dB. Remarkably, the optimization of $\tau$ boosts the performance of the new algorithm in the medium SNR regime. The impact of the mismatch statistics is almost insignificant for the proposed algorithm while for the LMMSE estimator the decrease in performance is up to one order of magnitude in the NMSE. Also, the BSVT recovery outperforms the SVT algorithm in all sampling regimes. It is also worth noting that for both the LMMSE and the

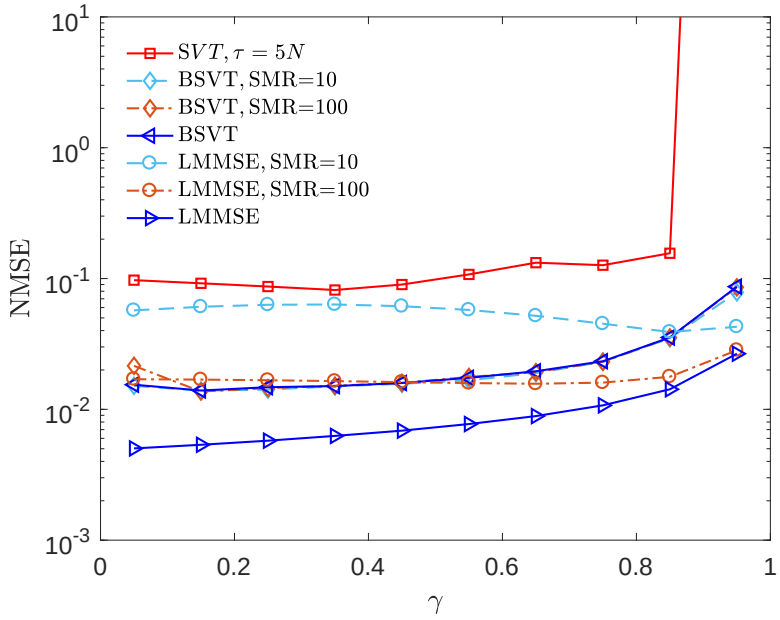



Figure 4.7. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT, LMMSE estimation and BSVT for different levels of mismatch, when SNR = 10 dB.

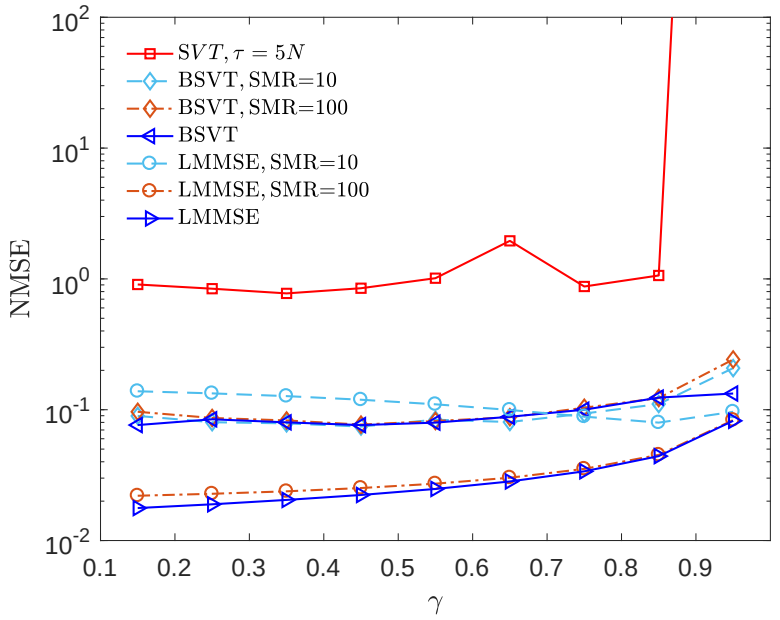



Figure 4.8. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, using SVT, LMMSE estimation and BSVT for different levels of mismatch, when SNR = 0 dB.

BSVT recovery the impact of mismatched statistics is less significant in the medium SNR regime compared with the high SNR regime. However, the robustness of the new approach holds regardless of the noise regime.

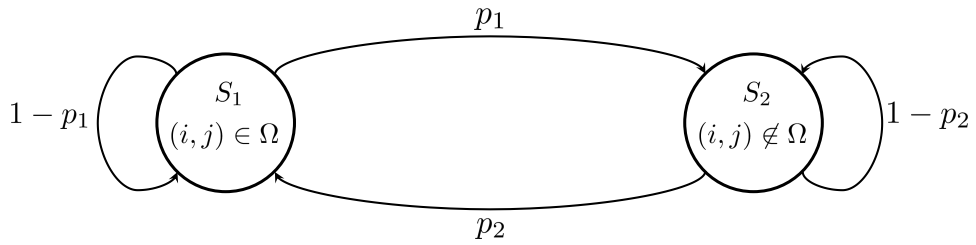



Figure 4.9. State diagram for the Markovian sampling model.

Figure 4.8 shows the performance of the SVT, BSVT, and LMMSE estimator for different levels of mismatch in the low SNR regime. Note that the case in which $\gamma = 0.05$ is not included in Figure 4.8 because neither SVT nor BSVT converge to a solution in the experiments carried out. However, for $\gamma \geq 0.15$ the new algorithm outperforms SVT by an order of magnitude regardless of the mismatch regime. In comparison with the LMMSE recovery, the proposed approach is again, robust to imperfect prior knowledge.

In practical scenarios, when the mismatch and noise regimes are difficult to assess, the choice between LMMSE and SVT is difficult to make. However, the BSVT algorithm provides robust recovery across all mismatch regimes for various noise levels. In view of this, the proposed recovery methods is a better alternative to state-of-the-art recovery methods.

It is worth noting that the impact of imperfect second order statistics for both LMMSE and BSVT is more significant for low levels of noise. Still, the change in recovery performance is less significant for the new algorithm in comparison with the LMMSE estimator.

Numerical results in this section assume that the locations of the missing entries are uniformly distributed. In practice, the set of missing entries might not satisfy this assumption. To account for structure in the observation pattern, the following section introduces a non-uniform sampling method.

### 4.4.4 Performance with non-uniform sampling

This section addresses the problem of recovering missing data when the subset of missing entries is not uniformly sampled. In practical scenarios, a sensor failure or a downtime in the communication line provides the operator with a number of consecutive unavailable measurements in the state variable vectors. Let $L_0$ be the number of consecutive missing entries. The expected value of $L_0$ varies depending on the reliability of the sensing infrastructure. In the uniform sampling model this scenario is not possible, and thus, a more general sampling procedure is introduced.

The proposed sampling model is based on a two-state Markov chain. In this setting, for each entry $(\mathbf{M})_{i,j}$ of the matrix $\mathbf{M}$, the finite state machine depicted

in Figure 4.9, is either in state $S_1$ in which case the entry $(i, j)$ is available to the operator, or in state $S_2$ in which case the entry is not available. As before, the set $\Omega$ contains all the entries from the matrix $\mathbf{M}$ that are available to the operator. In Figure 4.9, $p_1$ is the transition probability from state $S_1$ to $S_2$ and $p_2$ is the transition probability from $S_2$ to $S_1$. Hence, the expected value of the proportion of missing entries is given by the steady state probability of being in $S_2$.

The steady state probabilities for the Markov chain in Figure 4.9 are the solutions of the following equation

$$\mathbf{w}\mathbf{\Pi} = \mathbf{w}, \tag{4.21}$$

where the steady state vector $\mathbf{w}$ is

$$\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}, \tag{4.22}$$

and $w_1$ is the steady state probability of being in state $S_1$ and $w_2$ is the steady state probability of being in state $S_2$. In equation (4.21), $\mathbf{\Pi}$ is the transition matrix and for the model depicted in Figure 4.9 has the following expression

$$\mathbf{\Pi} = \begin{bmatrix} 1 - p_1 & p_1 \\ p_2 & 1 - p_2 \end{bmatrix}. \tag{4.23}$$

Solving (4.21) leads to

$$\mathbf{w} = \begin{bmatrix} \dfrac{p_2}{p_1 + p_2} & \dfrac{p_1}{p_1 + p_2} \end{bmatrix}. \tag{4.24}$$

Consequently, the expected value of the proportion of missing entries for the Markov chain sampling model is given by

$$\mathbb{E}[\gamma] = \frac{p_1}{p_1 + p_2}. \tag{4.25}$$

The expected number of consecutive missing entries, i.e., $\mathbb{E}[L_0]$, can be computed using

$$\mathbb{E}[L_0] = \sum_{l=0}^{n} l w_2 (1 - p_2)^l. \tag{4.26}$$

Solving (4.26) for $n \to \infty$ and combining with (4.25) leads to

$$\mathbb{E}[L_0] = \frac{1 - \mathbb{E}[\gamma]}{p_2^2}(1 - p_2). \tag{4.27}$$

Therefore, for any given $\gamma$ and $L_0$, using (4.25) and (4.27), $p_1$ and $p_2$ are identified such that on average the sampling model in Figure 4.9 has a proportion of missing entries $\gamma$ and the length of the vectors with consecutive missing entries $L_0$. Note that the case $\mathbb{E}[L_0] = 1$ reduces to the uniform sampling model with probability

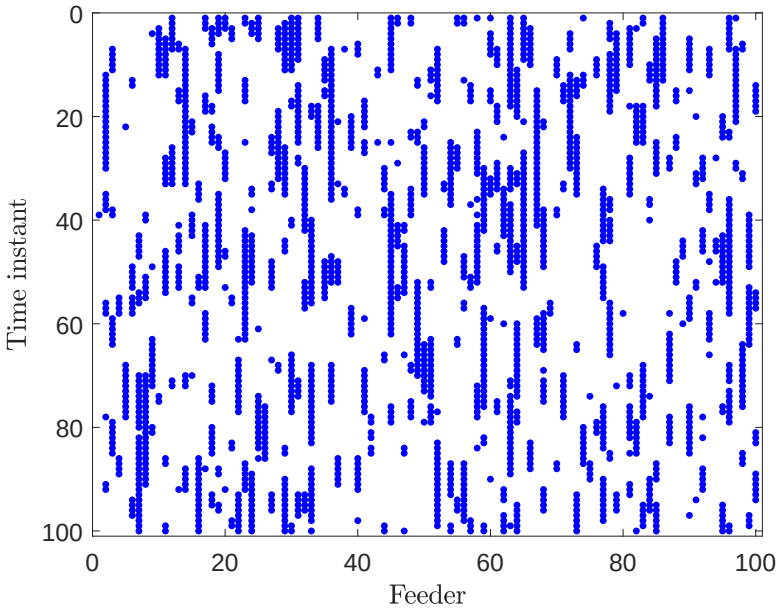



Figure 4.10. Positions of the observed entries, $\Omega$, generated by the Markovian model for a $100 \times 100$ matrix, when $\mathbb{E}[L_0] = M$ and $\mathbb{E}[\gamma] = 0.8$.

$\Pr[(i,j) \in \Omega] = 1 - \gamma$. In this framework, a comparison between the SVT and the BSVT-based recoveries is presented for the case in which the sampling pattern is not uniform. With the aim of considering the case in which a particular feeder does not provide any measurements, the expected length of the vectors with missing data is selected to be equal to the length of the state variable vectors, i.e., $\mathbb{E}[L_0] = M$. Figure 4.10 shows an example of a sampling pattern generated by the Markov-chain-based model, when $\mathbb{E}[L_0] = M$ and $\mathbb{E}[\gamma] = 0.8$.

Figure 4.11 compares the performance of the SVT-based recovery with the BSVT-based recovery for the case in which the matrix $\mathbf{M}$ is sampled using the Markov-chain-based sampling model with $\mathbb{E}[L_0] = M$. Different levels of mismatch are introduced to assess the robustness of the new algorithm to mismatched prior knowledge when the sampling pattern is not uniform. Remarkably, the performance of the proposed approach is not significantly affected by the amount of prior knowledge in any of the missing data regimes. Moreover, BSVT performs better than SVT when the sampling pattern is not uniform. A significant gain in performance is observed for small values of $\gamma$. Consider the following example for the sake of discussion, for a fixed tolerance of $10^{-2}$ in NMSE, the SVT algorithm recovers up to 4% of the entries of the matrix $\mathbf{M}$ while BSVT recovers 40% (See Figure 4.11). The improvement in the data recovering performance for the same level of tolerance is significant. Numerical results in this section show that BSVT is not only providing better performance than SVT when the entries are not uniformly sampled but it is also robust to mismatched statistics. The robustness of the new algorithm extends to

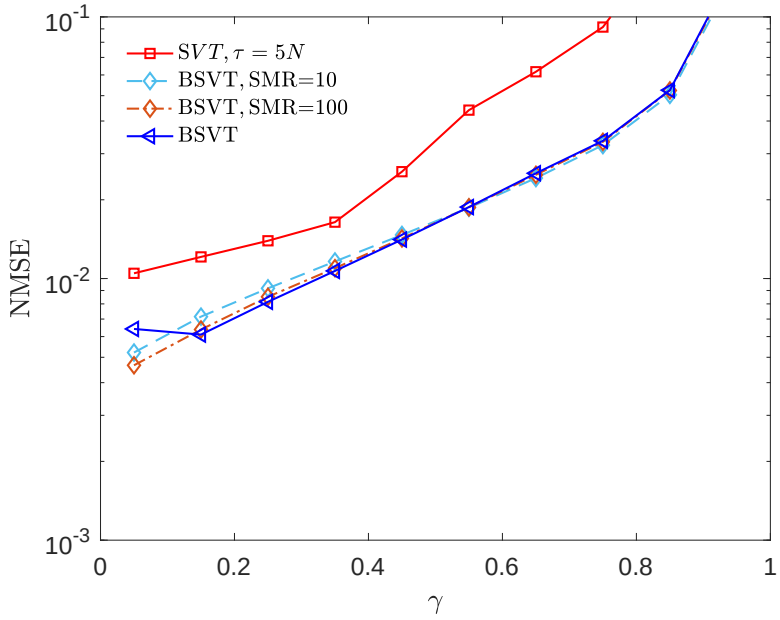



Figure 4.11. LV data recovery error measured by NMSE versus the proportion of missing entries, i.e., $\gamma$, for the Markov-chain-based sampling model, using SVT and BSVT for different levels of mismatch, when $\mathbb{E}[L_0] = M$ and SNR = 20 dB.

different sampling patterns. In view of this, BSVT represents a better alternative for recovering missing LV data in practical scenarios than SVT and LMMSE estimation.

## 4.5 Conclusions

This chapter introduced BSVT, a novel algorithm for recovering missing data in datasets that admit a low rank description. The proposed approach, combines the low computational cost of SVT with the optimality of the LMMSE estimator when the data source is modelled as a multivariate Gaussian random process and second order statistics are available. BSVT addresses some of the limitations of the SVT and LMMSE-based recoveries. In respect to the SVT algorithm the new approach addresses the issue of choosing the value of $\tau$ by calculating the optimal threshold at each iteration. As a consequence, the number of iterations required for the proposed algorithm is fewer than that of the SVT algorithm. Compared with the standard LMMSE estimator, BSVT is robust to inaccurate second order statistics. Interestingly, the impact of mismatch in the second order statistics across different noise regimes is similar for both LMMSE and BSVT. Specifically, for the same level of mismatch the performance degradation is more significant in the almost noiseless recovery for both estimation methods and decreases with the value of SNR. However, the change in performance for the BSVT algorithm is less significant compared to the LMMSE estimator.

The robustness of the new algorithm on both mismatched statistics and non-uniform sampling patterns is demonstrated through numerical simulations. The gain in performance compared to SVT is noticeable for both uniform and non-uniform sampling models regardless of the mismatch introduced in the second order statistics. Ultimately, the proposed algorithm is shown to provide a robust and low complexity method to recover missing data in high-dimensional datasets that follow a multivariate Gaussian distribution.

# Chapter 5

# Joint recovery of missing data in multiple datasets

## 5.1 Introduction

An urban system contains a set of interacting subsystems such as the traffic monitoring system, the air quality monitoring system or the low voltage distribution system. The interdependencies between the different components of the urban system give rise to correlation between the dataset produced by each subsystem.

The analysis presented in Chapter 3 and Chapter 4 pertains to the independent recovery of the missing data from each subsystem. However, when the recovery of missing data is performed independently for each dataset the interaction between the urban subsystems is not captured. In the following, a joint estimation framework is proposed to enable the use of the additional type of correlation in a missing data recovery setting.

As an example consider the problem of recovering missing data collected by the traffic monitoring system. Due to issues with the sensing infrastructure, the number of observations available is usually not sufficient for recovery in practical settings. In the joint recovery case, the data collected by the pollution monitoring system is used to enable the recovery of missing data in the traffic monitoring system. This is achieved by combining the traffic and air quality data matrices into a single matrix. As a result, the combined matrix contains sufficient entries to facilitate the recovery using matrix completion-based techniques.

It is worth noting that in the joint recovery case, the are two types of correlation between the entries of the combined matrix. First, the intra-correlation that refers to the correlation between the entries within each dataset. This is the type of correlation that is exploited in the independent recovery scenario. Second, the cross-correlation is defined as the correlation between the data points from two different datasets. In this example considered above, this is the correlation between the traffic

level and the air quality data. In contrast to the independent recovery case, the joint recovery case exploits both types of correlation. Interestingly, exploiting the cross-correlation between datasets facilitates the recovery of both datasets when the number of available entries is not sufficient for independent recovery.

In the following, Section 5.2 introduces the system model and the fundamental limits of the joint recovery problem using two datasets are presented in Section 5.3. In particular, the conditions in which the joint recovery of two datasets requires fewer entries compared to the independent case are identified. Based on the fundamental limits, Section 5.4 presents a lower bound for the probability of recovery when the locations of the missing entries are uniformly distributed in each dataset. Furthermore, in Section 5.5 the bounds are used to minimize the sampling complexity for the recovery of two datasets. A numerical analysis for joint recovery performance using the SVT and BSVT algorithms is presented in Section 5.6. Finally, the conclusions are presented in Section 5.7.

## 5.2 System model

Consider two sensing infrastructures that aggregate observations in the matrices $\mathbf{M}_1 \in \mathbb{R}^{M\times N}$ and $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$, respectively. To simplify the analysis, both matrices are of the same size. Following with the introduction example, $\mathbf{M}_1$ contains traffic flow data and $\mathbf{M}_2$ contains measurements of the $CO_2$ level over the same section of the road. When the data from each sensing infrastructure is correlated the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are low rank or approximately low rank. Without loss of generality, we assume that both matrices are low rank. The case in which both are approximately low rank is treated as in Section 3.2. Therefore, the following holds

$$\operatorname{rank}(\mathbf{M}_1) = r_1, \tag{5.1}$$

and

$$\operatorname{rank}(\mathbf{M}_2) = r_2. \tag{5.2}$$

Let us aggregate the data from $\mathbf{M}_1$ and $\mathbf{M}_2$ such that the resulting matrix $\mathbf{M} \in \mathbb{R}^{2M\times N}$ is given by

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix}, \tag{5.3}$$

and denote the rank of the resulting matrix by

$$\operatorname{rank}(\mathbf{M}) = r. \tag{5.4}$$

Consequently, the following inequalities hold

$$\begin{aligned} 1 \leq & r_1 \leq \min\{M, N\} \\ 1 \leq & r_2 \leq \min\{M, N\}. \end{aligned} \tag{5.5}$$

Within this setting, the following section derives for the first time theoretical bounds for the joint estimation of missing entries in the two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ and establishes theoretical guarantees. The joint recovery of the two matrices is said to be beneficial when the minimum number of observations needed to recover the missing entries of the combined matrix is fewer than the number of entries required to perform independent recovery of each matrix. To that end, the following section provides upper and lower bounds on the rank of the combined matrix, i.e. rank($\mathbf{M}$), as a function of the rank of the two matrices. Rank conditions that ensure the recovery of the matrix $\mathbf{M}$ is beneficial compared to the independent recovery of the two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are derived.

## 5.3 Rank conditions for the joint recovery to be beneficial

This section investigates the theoretical guarantees for the joint recovery of two matrices to be beneficial, that is when it requires fewer entries than what is necessary to perform the independent recovery of each matrix. In general, the minimum number of entries required for the recovery of one dataset is a function of the rank and the size of the matrix. The rank of the matrix varies depending on the level of correlation between the measurements contained in the matrix. In the following, the correlation between measurements from the same dataset is referred to as intra-correlation. In this setting, the rank of the matrix and the number of entries required for the independent recovery of one dataset depends on the level of intra-correlation between the state variables contained in the dataset. In the joint recovery case another type of correlation arises. The term cross-correlation refers to the correlation between the two datasets combined. As an example consider the case in which the matrix $\mathbf{M}_1$ contains traffic flow density data and the matrix $\mathbf{M}_2$ contains $CO_2$ emission measurements. In addition to the correlation between state variables within the same dataset, there exists a correlation between the entries in $\mathbf{M}_1$ and the entries in $\mathbf{M}_2$. In this particular example this is caused by the fact that car emissions are one of the main sources of $CO_2$. A larger number of cars results in a higher concentration of $CO_2$ and vice versa.

The combination of the datasets into a single matrix increases the size of the matrix, and therefore, the joint recovery is beneficial depending on the trade-off

between the size and the rank of the combined matrix. Note that the rank of the combined matrix depends on both the intra and the cross-correlation. In order to benefit from the joint recovery the rank of the combined matrix needs to compensate for the increase on the matrix size.

In the following, rank conditions for benefiting from the joint recovery of two datasets are characterized. To that end, upper and lower bounds on the rank of the combined matrix are provided in the following section.

### 5.3.1 Bounds on the rank of the combined matrix

This section assesses the impact of both intra and cross-correlation on the rank of the combined matrix. The following lemma provides lower and upper bounds for the rank of the combined matrix based on the individual rank of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$.

**Lemma 1.** *Let* $\mathbf{M}_1 \in \mathbb{R}^{M\times N}$ *and* $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$. *Define the combined matrix* $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2M\times N}$. *Then, the following holds:*

$$\max\{\text{rank}(\mathbf{M}_1), \text{rank}(\mathbf{M}_2)\} \leq \text{rank}(\mathbf{M}) \leq \text{rank}(\mathbf{M}_1) + \text{rank}(\mathbf{M}_2). \tag{5.6}$$

*Proof.* See Appendix A. □

The intra-correlation determines the rank of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ which define the minimum and maximum possible value for $r$. A smaller value of intra-correlation in one of the datasets results in an increase on the lower bound for $r$. On the other hand, the cross-correlation governs the value of $r$ within the limits defined by Lemma 1. Indeed, a larger value of cross-correlation results in a value of $r$ that is closer to the lower bound while a smaller value of cross-correlation generates a combined matrix with a rank that is close to the upper bound. In other words, the intra-correlation defines the limits of $r$ and the cross-correlation governs the value of $r$ within the limits.

A description of the bounds for the rank of the combined matrix in Lemma 1, in terms of the intersection of the row subspaces of the two matrices follows. The rank of the matrix $\mathbf{M}$ is determined by (3.23(a) in [100])

$$r = r_1 + r_2 - d, \tag{5.7}$$

where $d$ is the dimension of the intersection of the row subspaces of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ given by

$$d = \dim[\mathcal{C}(\mathbf{M}_1^T) \cap \mathcal{C}(\mathbf{M}_2^T)], \tag{5.8}$$

where $\mathcal{C}(\mathbf{M}_1^T)$ denotes the column subspace of the matrix $\mathbf{M}_1^T$. Consequently, the rank of the combined matrix is equal to $r^*$ when the dimension of the intersection of the row subspaces is [100]

$$\dim[\mathcal{C}(\mathbf{M}_1^T) \cap \mathcal{C}(\mathbf{M}_2^T)] = r_1 + r_2 - r^*. \tag{5.9}$$

A larger dimension of the intersection of the two row subspaces results in a smaller rank for the combined matrix. In contrast, a larger value of $r$ is obtained when the dimension of the intersection of the row subspaces is smaller, and similarly, the same effect on the rank is observed for the level of cross-correlation. In particular, a higher level of cross-correlation results in a lower combined rank and vice versa. Therefore, in the context of determining the rank of the combined matrix, the dimension of the intersection of the row subspaces exhibits a similar effect to the level of cross-correlation.

In the matrix completion-based recovery setting, the minimum number of entries required depends on the rank of the matrix [35] and [59]. In the following section, theoretical guarantees for the joint recovery of two matrices to be beneficial are characterized. Specifically, necessary and sufficient conditions are derived such that the joint recovery requires fewer entries than in the independent case.

### 5.3.2 Theoretical guarantees for the joint recovery of two datasets to be beneficial

Lemma 1 shows the relationship between the two types of correlation and the rank of the combined matrix. The following analyzes the link between the rank of the combined matrix and the recovery guarantees in a matrix completion framework. In particular, a necessary condition depending on the rank and the size of the combined matrices and a sufficient condition depending on the rank of the combined matrix are presented. Both results are stated in Theorem 2 which provides a necessary and sufficient condition for the joint recovery of two datasets to be beneficial.

To facilitate the analysis for the joint recovery of two datasets, let us recall the theoretical limits for the recovery of one dataset presented in [59]. To this end, the following definitions are necessary [59]:

**Definition 5.** *Given a random matrix* $\mathbf{X} \in \mathbb{R}^{M \times N}$ *of arbitrary distribution, an* $(M \times N, k)$ *code consists of*

1. *linear observations* $(\langle \mathbf{A}_1, \cdot \rangle, \cdots, \langle \mathbf{A}_k, \cdot \rangle)^T : \mathbb{R}^{M \times N} \to \mathbb{R}^k$*;*
2. *a measurable decoder* $g : \mathbb{R}^k \to \mathbb{R}^{M \times N}$*.*

*For given observation matrices* $\mathbf{A}_i$*, the decoder* $g$ *achieves error probability* $\varepsilon$ *if*

$$\Pr[g\Big((\langle \mathbf{A}_1, \cdot\rangle, \cdots, \langle \mathbf{A}_k, \cdot\rangle)^T\Big) \neq \mathbf{X}] \leq \varepsilon.$$

**Definition 6.** *For* $\varepsilon \geq 0$*,* $\mathcal{S} \subseteq \mathbb{R}^{M\times N}$ *is an* $\varepsilon$*-support set of the random matrix* $\mathbf{X} \in \mathbb{R}^{M\times N}$ *if it is a nonempty bounded set and* $\Pr[\mathbf{X} \in \mathcal{S}] \leq 1 - \varepsilon$*.*

In view of this, Remark 2 in [59] states that given a random matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ with $\text{rank}(\mathbf{X}) = r$, for every $\varepsilon > 0$, there exists a decoder that achieves error probability $\varepsilon$ for almost all Lebesgue observation matrices $\mathbf{A}_i$, $i \in \{1, \cdots, k\}$, if $k > (M + N - r)r$. Moreover, the same result holds when the matrices $\mathbf{A}_i$ are rank-one observation matrices with $i \in \{1, \cdots, k\}$.

Using the analysis in [59] the low rank matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ can be successfully recovered when the number of observations for the first matrix satisfies

$$k_1 > (M + N - r_1)r_1, \tag{5.10}$$

and the number of available observations for the second matrix obeys

$$k_2 > (M + N - r_2)r_2. \tag{5.11}$$

This result is based on the assumption that for the random matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ there exist a $\sigma$-measure $\mu_{\mathbf{M}_1}$ and a $\sigma$-measure $\mu_{\mathbf{M}_2}$ and that both measures admit a Lebesgue decomposition. For the combined matrix $\mathbf{M}$, the $\sigma$-measure is obtained as the product of the measures of $\mathbf{M}_1$ and $\mathbf{M}_2$ [110]

$$\mu_{\mathbf{M}} = \mu_{\mathbf{M}_1} \times \mu_{\mathbf{M}_2}. \tag{5.12}$$

Moreover, since $\mu_{\mathbf{M}}$ is a $\sigma$-measure it also admits a Lebesgue decomposition [111] and [112]. Hence, the result in [59] applies for the combined matrix $\mathbf{M}$ without any additional assumptions

$$k > (2M + N - r)r, \tag{5.13}$$

where $k$ is the number of the available observations for the joint recovery case. Note that in this framework, the number of observations for the matrix $\mathbf{M}$ is equal to

$$k = k_1 + k_2. \tag{5.14}$$

Therefore, the inequality in (5.13) is equivalent to

$$k_1 + k_2 > (2M + N - r)r. \tag{5.15}$$

It is also worth noting that the recovery guarantees presented in [59] and extended here to a joint recovery framework, are achievability results using linear observations where the recovery method is not known. Denote the lower bound on the number of observations required for the recovery of the matrix $\mathbf{M}_1$ by $\underline{k}_1$. Therefore, $\underline{k}_1$ is given by

$$\underline{k}_1 = (M + N - r_1)r_1. \tag{5.16}$$

Similarly, denote by $\underline{k}_2$ the lower bound on the number of observations required for the recovery of the matrix $\mathbf{M}_2$, then

$$\underline{k}_2 = (M + N - r_2)r_2. \tag{5.17}$$

Moreover, for the joint recovery, the lower bound on the number of observations required is given by

$$\underline{k} = (2M + N - r)r. \tag{5.18}$$

In view of this, the joint recovery of two datasets requires fewer observations compared to the independent recovery case when

$$\underline{k} < \underline{k}_1 + \underline{k}_2. \tag{5.19}$$

The joint recovery of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ is beneficial when the recovery of $\mathbf{M}$ requires a smaller number of observations in comparison with the independent recovery of both matrices. When $\mathbf{M}_1$ and $\mathbf{M}_2$ are recovered independently, the minimum number of observations is given by the sum of $\underline{k}_1$ and $\underline{k}_2$.

Figure 5.1 depicts a graphical representation of the inequalities in (5.10), (5.11) and (5.13) that describe the lower bound on the number of entries required to recover the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$, respectively. When the joint recovery is beneficial, the three inequalities divide the $(k_1, k_2) \in \mathbb{N}^2$ plane in seven regions.

$\mathcal{R}_1$ is given by

$$\mathcal{R}_1 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 \leq \underline{k}_1,\, k_2 \leq \underline{k}_2,\, k_1 + k_2 \leq \underline{k}\Big\}, \tag{5.20}$$

and contains all the pairs ($k_1$,$k_2$) for which the recovery of the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ is not feasible. In other words, none of the inequalities is satisfied which means that there are not enough observations to recovery either of the three matrices. Following with the introduction example, $\mathcal{R}_1$ corresponds to the case in which the number of observations from both the traffic monitoring system and the air quality monitoring system does not suffice for either the independent nor the joint recovery of the two datasets.

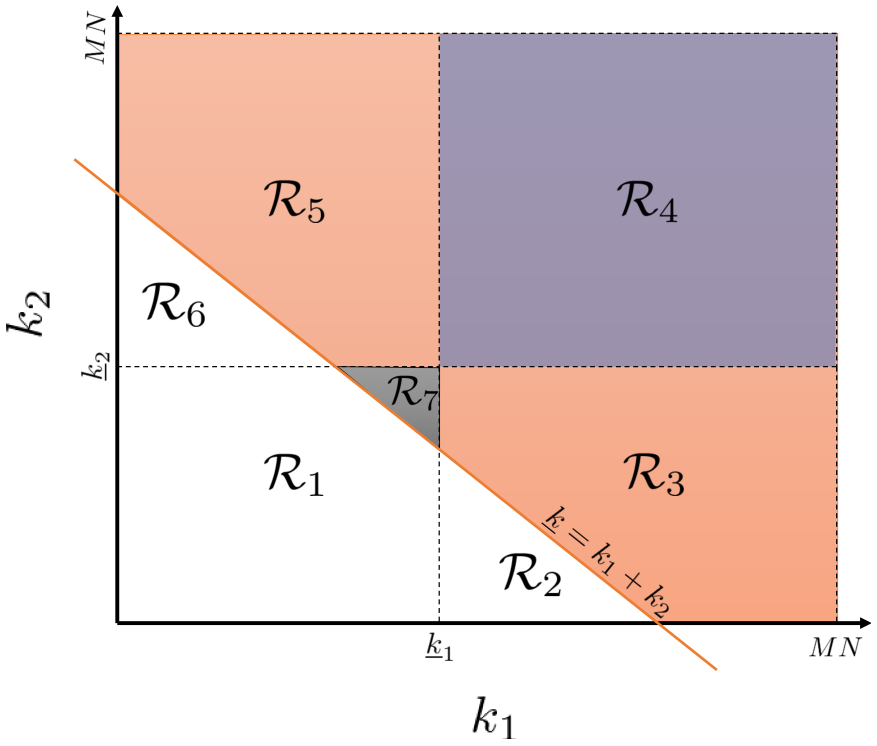



Figure 5.1. Recovery regions for the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$, for a large value of cross-correlation when $r_2/r_1$ is small.

The second region, $\mathcal{R}_2$ is given by

$$\mathcal{R}_2 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 > \underline{k}_1,\ k_2 \leq \underline{k}_2,\ k_1 + k_2 \leq \underline{k}\Big\}, \tag{5.21}$$

and contains all the pairs ($k_1$,$k_2$) for which the matrix $\mathbf{M}_1$ is recovered but the recovery of $\mathbf{M}_2$ and $\mathbf{M}$ is not feasible. Precisely, only the inequality in (5.10) is satisfied and the other two are not. In the introduction example, $\mathcal{R}_2$ is equivalent to the case in which the number of traffic flow observations is sufficient for the independent recovery of the data collected from the traffic monitoring system. However, because the number of $CO_2$ level observations is small, both the independent recovery of the data from the air quality monitoring system and the joint recovery of the two datasets are not feasible.

Region $\mathcal{R}_3$ is defined by

$$\mathcal{R}_3 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 > \underline{k}_1,\ k_2 \leq \underline{k}_2,\ k_1 + k_2 > \underline{k}\Big\}, \tag{5.22}$$

and contains all the pairs ($k_1$,$k_2$) for which the matrices $\mathbf{M}_1$ and $\mathbf{M}$ are recovered but the recovery of $\mathbf{M}_2$ is not feasible. In this case, the joint recovery is possible because there is a large number of available entries in $\mathbf{M}_1$. In other words, the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ allows the recovery of the second matrix with a limited number of observations by exploiting the larger number of observations

from the first matrix. Coming back to the example presented in the introduction, $\mathcal{R}_3$ corresponds to the case in which the number of traffic flow observations is sufficient for the independent recovery of the data collected from the traffic monitoring system but the number of $CO_2$ level observations is not enough for the independent recovery of the data from the air quality monitoring system. However, the cross-correlation between the two datasets facilitates the recovery of both types of data in a joint recovery framework. Specifically, the total number of observations is sufficient for the recovery of the combined matrix $\mathbf{M}$.

Region $\mathcal{R}_4$ is given by

$$\mathcal{R}_4 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 > \underline{k}_1,\, k_2 > \underline{k}_2,\, k_1 + k_2 > \underline{k}\Big\}, \tag{5.23}$$

and contains all the pairs ($k_1$,$k_2$) for which all three matrices are recovered. The number of observations is large enough for both the independent and the joint recovery. Precisely, all three inequalities (5.10), (5.11) and (5.13) are satisfied. Following with the introduction example, $\mathcal{R}_4$ resembles the case in which the number of observations from both the traffic monitoring system and the air quality monitoring system is sufficient for both the independent and the joint recovery of the two types of data.

Region $\mathcal{R}_5$ is defined by

$$\mathcal{R}_5 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 \leq \underline{k}_1,\, k_2 > \underline{k}_2,\, k_1 + k_2 > \underline{k}\Big\}, \tag{5.24}$$

and contains all the pairs ($k_1$,$k_2$) for which the matrices $\mathbf{M}_2$ and $\mathbf{M}$ are recovered but the recovery of $\mathbf{M}_1$ is not feasible. In this case, the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ allows the recovery of the first matrix with a limited number of observations by exploiting the larger number of observations from the second matrix. In the example presented in the introduction section, $\mathcal{R}_5$ corresponds to the case in which the number of observations from the traffic monitoring system is not enough for the independent recovery of the traffic data. However, the number of observations from the air quality monitoring system is large enough that it facilitates both the independent recovery of the $CO_2$ level data and the joint recovery of both datasets. Similar to the region $\mathcal{R}_3$, in this case a larger number of observations from one of the datasets enables the joint recovery of both types of data. In contrast, in the independent recovery paradigm, only the dataset with the larger number of observations could be recovered.

Region $\mathcal{R}_6$ is given by

$$\mathcal{R}_6 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 \leq \underline{k}_1,\, k_2 > \underline{k}_2,\, k_1 + k_2 \leq \underline{k}\Big\}, \tag{5.25}$$

and contains all the pairs ($k_1$,$k_2$) for which the matrix $\mathbf{M}_2$ is recovered and the recovery of $\mathbf{M}$ and $\mathbf{M}_1$ is not feasible. In this case, only the inequality in (5.11) is satisfied while the other two are not. Coming back to the example presented in the introduction, $\mathcal{R}_6$ resembles the case in which the number of traffic flow observations is not sufficient for the independent recovery of the traffic data. On the other hand, the air quality monitoring system collected enough observations to facilitate the recovery of $CO_2$ level data. However, the joint recovery is not possible because the total number of observations from both datasets is not sufficient.

Finally, Region $\mathcal{R}_7$ is given by

$$\mathcal{R}_7 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 \leq \underline{k}_1,\, k_2 \leq \underline{k}_2,\, k_1 + k_2 > \underline{k}\Big\}, \tag{5.26}$$

and contains all the pairs ($k_1$,$k_2$) for which the matrix $\mathbf{M}$ is recovered and the independent recovery of $\mathbf{M}_1$ and $\mathbf{M}_2$ is not feasible. In other words, the independent recovery of the two datasets is not possible but by combining them in a joint matrix, the cross-correlation facilitates the recovery of both types of data in a joint recovery paradigm. In the introduction example, $\mathcal{R}_7$ is equivalent to the case in which the number of observations from the traffic monitoring system is not enough for the recovery of the traffic flow data. Similarly, the number of data points collected by the air quality monitoring system does not suffice for the independent recovery of the $CO_2$ level data. However, in a joint recovery context, the missing data from both datasets is successfully recovered. In addition, the existence of region $\mathcal{R}_7$ guarantees that the joint recovery is beneficial for some pairs ($k_1$,$k_2$). In the following, the theoretical guarantees for the joint recovery of two datasets to be beneficial are obtained by verifying the existence of region $\mathcal{R}_7$.

Coming back to the example presented in the introduction, the traffic flow data and the $CO_2$ level data can be independently recovered only when the pair ($k_1$,$k_2$) satisfies the conditions of region $\mathcal{R}_4$. On the other hand, in a joint recovery framework the data from both the traffic monitoring system and the air quality system is recovered when the pair ($k_1$,$k_2$) is located in region $\mathcal{R}_3$, $\mathcal{R}_4$, $\mathcal{R}_5$ or $\mathcal{R}_7$.

Moreover, the existence of region $\mathcal{R}_7$ shows that there are cases in which neither of the two datasets is recovered independently but the joint recovery is viable. Motivated by this observation the following definition captures the notion of *beneficial joint recovery*

**Definition 7.** *The joint recovery of two matrices,* $\mathbf{M}_1$, $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$ *of rank* $r_1$ *and* $r_2$*, respectively, is beneficial if there exists* $(k_1, k_2) \in \mathcal{R}_7$ *given by*

$$\begin{aligned}\mathcal{R}_7 = \Big\{(k_1, k_2) \in \mathbb{N}^2 : k_1 \leq (M + N - r_1)r_1,\, k_2 \leq (M + N - r_2)r_2,\\ k_1 + k_2 > (2M + N - r)r\Big\}.\end{aligned} \tag{5.27}$$

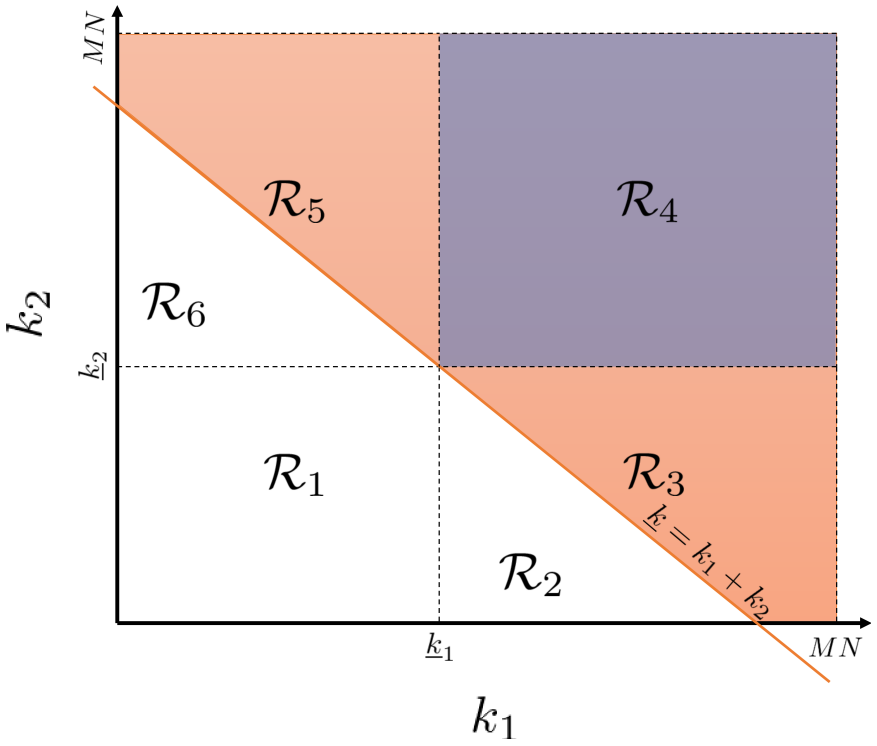



Figure 5.2. Recovery regions for the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ for a moderate value of cross-correlation when $r_2/r_1$ is small.

Is it worth noting that the geometry of the recovery regions changes depending on the rank of the matrices. Specifically, the size and shape of the regions $\mathcal{R}_1$ and $\mathcal{R}_4$ depend on $\underline{k}_1$ and $\underline{k}_2$ which are a function of the rank and the size of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$. On the other hand, the size and shape of the regions $\mathcal{R}_2$, $\mathcal{R}_3$, $\mathcal{R}_5$ $\mathcal{R}_6$ and $\mathcal{R}_7$ vary with the value of $\underline{k}$ which ultimately depends on the matrix $\mathbf{M}$. Since the matrix $\mathbf{M}$ is obtained by combining $\mathbf{M}_1$ and $\mathbf{M}_2$, the level of correlation between the two datasets dictates the rank and ultimately the minimum number of observations required for the recovery of the combined matrix. In this context, for different values of correlation the geometry of the recovery regions is changing.

Figure 5.2 depicts the theoretical bounds for the minimum number of entries required for independent and joint recovery for datasets with a moderate level of cross-correlation. In contrast to Figure 5.1, the rank of the combined matrix, i.e. $r$, is larger which results in a larger number of entries required for the joint recovery. Therefore, for a particular value of $r$ the intersection between the line described by $\underline{k} = k_1 + k_2$ and $\mathcal{R}_4$ is a single point. This value of $r$ represents the minimum value for which the set defined by the region seven is empty. In other words, the joint recovery is beneficial when the rank of the combined matrix is smaller than the $r$ for which the intersection between the line $\underline{k} = k_1 + k_2$ and $\mathcal{R}_4$ is a single point. On the other hand, for larger values of $r$, Figure 5.3 shows that the intersection between

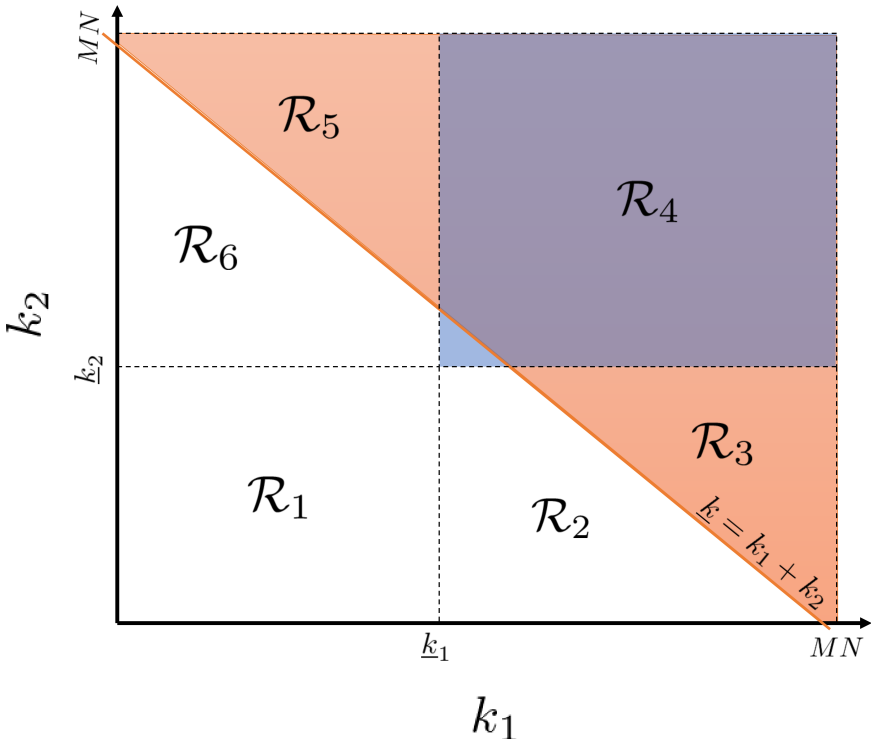



Figure 5.3. Recovery regions for the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ for a small value of cross-correlation when $r_2/r_1$ is small.

the boundary defined by $\underline{k} = k_1 + k_2$ and $\mathcal{R}_4$ enlarges beyond a single point and $\mathcal{R}_7$ becomes an empty set.

Based on the above discussion, a necessary condition for the joint recovery to be beneficial is that $\underline{k} = k_1 + k_2$ and $\mathcal{R}_4$ do not intersect. However, $\mathcal{R}_4$ is determined by the rank of the two matrices. In contrast, the line $\underline{k} = k_1 + k_2$ depends on the rank of the combined matrix. As a result of Lemma 1, it follows that the value of $r$ is bounded by

$$\max\{r_1, r_2\} \leq r \leq r_1 + r_2. \tag{5.28}$$

In addition, for large values of cross-correlation, the value of $r$ is closer to the lower bound while for smaller values of cross-correlation the value of $r$ is closer to the upper bound. As expected, the graphical representations in Figure 5.1 and 5.3 show that the joint recovery is beneficial for large values of cross-correlation. However, there exists cases in which a large value of cross-correlation does not guarantee a joint recovery that is advantageous.

Figure 5.4 depicts the same set of inequalities given by (5.10), (5.11) and (5.13) when the value of cross-correlation is large and the ratio between the rank of the two datasets is large. Interestingly, Figure 5.4 shows that the conditions under which the joint recovery is beneficial depend on both the value of cross-correlation and the ratio between $r_2$ and $r_1$. This is due to the fact that the lower bound on the rank of the combined matrix is given by the maximum rank. In this case, the maximum rank

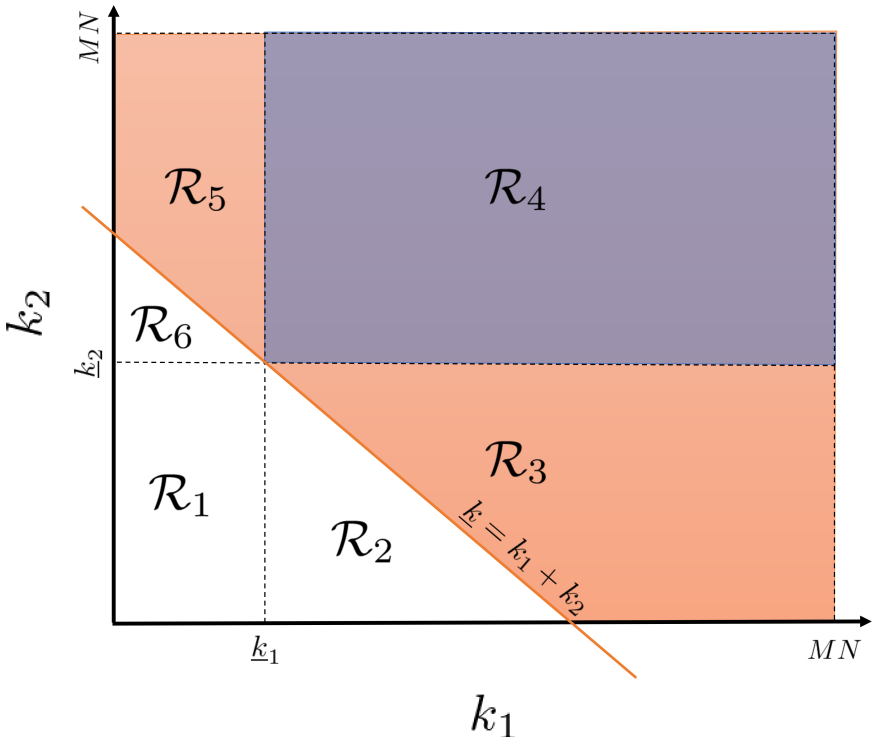



Figure 5.4. Recovery regions for the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ for a large value of cross-correlation when $r_2/r_1$ is large.

becomes the governing factor in the joint recovery setting and does not allow the recovery using fewer entries compared to the independent case. Therefore, when the ratio between $r_2$ and $r_1$ is large, the line described by $\underline{k} = k_1 + k_2$ and $\mathcal{R}_4$ intersect in at least one point which is equivalent to the joint recovery not being beneficial regardless of the rank of the combined matrix. In this context, the following lemma provides a condition for the joint recovery to be beneficial when the value of $r$ is minimized. When the condition in Lemma 2 is not satisfied, the joint recovery is not advantageous.

**Lemma 2.** *(Necessary condition for the joint recovery to be beneficial) Let* $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M\times N}$ *with rank* $r_1$ *and* $r_2$ *respectively, there exists a value of* $r$ *for which the joint recovery is beneficial if*

$$M\Big(\min\{r_1, r_2\} - \max\{r_1, r_2\}\Big) > \min\{r_1, r_2\}\Big(\min\{r_1, r_2\} - N\Big). \tag{5.29}$$

*Proof.* See Appendix A. □

In other words, when the condition in Lemma 2 is not satisfied, regardless of the rank of the combined matrix, the joint recovery of the two datasets is not beneficial. On the other hand, when the condition in Lemma 2 is satisfied the value of the cross-correlation dictates if the joint recovery is beneficial. When the inequality in (5.29) is not satisfied, the ratio between $r_2$ and $r_1$ is too large and the joint recovery

is not beneficial. In the example presented in the introduction section, where the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ contain traffic flow and $CO_2$ level data, respectively, Lemma 2 provides a condition for verifying the compatibility of the two types of data in a joint recovery context. Specifically, based on the size and the rank of the traffic data and air quality data matrices, Lemma 2 evaluates the possibility benefiting from combining the traffic flow data and the $CO_2$ level data into a larger matrix. Based on the observation presented in Figure 5.4, when there is a considerable difference between the rank of the two matrices combined, the joint recovery is not beneficial. In view of this, the inequality in (5.29) verifies if the ratio between the rank of the traffic data matrix and the rank of the air quality data matrix is small enough to facilitate the possibility of benefiting from the joint recovery. When Lemma 2 is satisfied, the rank of the combined matrix dictates if the joint recovery is beneficial or not. It is also worth noting that the condition in Lemma 2 does not depend on the combined matrix $\mathbf{M}$. The following theorem provides an additional constraint on the rank of the combined matrix to guarantee that the joint recovery is beneficial.

**Theorem 2.** *(Necessary and sufficient condition for benefiting from the joint recovery) Let $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M\times N}$, with rank $r_1$ and $r_2$, the joint recovery of the two matrices requires fewer observations than the independent recovery if*

$$M\Big(\min\{r_1, r_2\} - \max\{r_1, r_2\}\Big) > \min\{r_1, r_2\}\Big(\min\{r_1, r_2\} - N\Big), \tag{5.30}$$

*and the rank of the combined matrix satisfies*

$$r < \frac{2M + N - \sqrt{(2M+N)^2 - 4(M+N)(r_1+r_2) + 4r_1^2 + 4r_2^2}}{2}. \tag{5.31}$$

*Proof.* See Appendix A. □

Note that the necessary condition from Theorem 2, that is (5.30), is the same as in Lemma 2 and that the inequality depends only on the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ and not on the combined matrix $\mathbf{M}$. In contrast, the sufficient condition in (5.31) provides an upper bound for the rank of the combined matrix such that the total number of observations required for the joint recovery is fewer compared to the independent recovery case. Theorem 2 provides a necessary and sufficient condition for the joint recovery of two matrices to be beneficial. In practice, the number of available entries depends on the sampling pattern considered. However, the achievability results presented in this section do not assume any sampling strategy. Typically, recovery guarantees for matrix completion assume that the locations of the missing entries are sampled uniformly at random from the set with all the entries [35], [36], [50] and [51]. In this context, the next section introduces a probabilistic framework in which the probabilities of independent and joint recovery are characterized for

an uniform sampling pattern. Although the applicability of uniform sampling in a realistic scenario is limited, the analysis is more tractable.

## 5.4 Recovery guarantees under random uniform sampling

In the following we address the joint recovery problem in a probabilistic framework. The probability of the recovery in the independent and joint case are characterized depending on the sampling structure considered. This is a particular case of the fundamental limits presented in Section 5.3.2. In the uniform sampling case considered in this section, the locations of the available observations are uniformly distributed in the matrix. In the following, the entries in $\mathbf{M}_1$ are sampled uniformly at random with probability $p_1$ and the entries in $\mathbf{M}_2$ are sampled uniformly at random with probability $p_2$.

In a practical setting the values of $p_1$ and $p_2$ depend on the reliability of the sensing infrastructure of the first and second subsystem, respectively. In the uniform sampling scenario, each sensor provides the same degree of reliability in the sensing infrastructure that is part of. In other words, the number of observations obtained from each sensor over a period of time is approximately the same for all sensors. Moreover, the observations from each sensor are uniformly distributed in the temporal dimension. Within this setting, the number of observations for the matrix $\mathbf{M}_i$ with $i \in \{1, 2\}$ follows a binomial distribution with parameters $MN$ and $p_i$.

The set of available entries for joint recovery is the union of the sets of available entries for $\mathbf{M}_1$ and $\mathbf{M}_2$. Note that for the combined matrix the sampling pattern is not uniform unless $p_1 = p_2$. The probability of sampling the matrix $\mathbf{M}$ is given by

$$p(j) = \begin{cases} p_1, & \text{if} \quad 1 \leq j \leq MN \\ p_2, & \text{if} \quad MN + 1 \leq j \leq 2MN. \end{cases} \tag{5.32}$$

In other words, the first $MN$ entries which correspond to the matrix $\mathbf{M}_1$ are sampled with probability $p_1$ and the rest of the entries corresponding to the matrix $\mathbf{M}_2$ are sampled with probability $p_2$. Consequently, the number of observations for the matrix $\mathbf{M}$ follows a Poisson binomial distribution with parameters $2MN$ and $p$.

However, the successful recovery of the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ depends on the number of available entries for each matrix. In this framework, the number of available entries for $\mathbf{M}_1$ and $\mathbf{M}_2$ is a random variable generated by a binomial distribution with parameters $MN$ and $p_1$, and $MN$ and $p_2$ respectively. Let us denote the number of available entries for the first dataset by $K_1$ and the number of available entries for the second dataset by $K_2$. Note that $K_1$ and $K_2$ are random

Table 5.1. Rank values considered in the numerical simulations and the fundamental limit.

| $r_1$ | $r_2$ | $r$ | Fundamental limit imposed by Theorem 2 |
|---|---|---|---|
| 6 | 6 | 6 | 9.05 |
| 6 | 6 | 9 | |
| 6 | 6 | 10 | |
| 6 | 6 | 12 | |
| 6 | 9 | 10 | 11.3 |
| 6 | 9 | 11 | |
| 6 | 9 | 12 | |
| 6 | 9 | 14 | |
| 6 | 18 | 19 | 17.78 |
| 6 | 18 | 22 | |

variables that depend on the size of the matrices and the sampling probabilities $p_1$ and $p_2$. In this context, the number of observations for the matrix $\mathbf{M}$ is the sum of the random variables $K_1$ and $K_2$. Moreover, the sum $K_1 + K_2$ is a random variable distributed as a Poisson binomial with parameters $2MN$ and $p$ [113], where $p$ is defined in (5.32).

In this context define by $E_1$ the event in which the matrix $\mathbf{M}_1$ is recovered. Therefore, $E_1$ is given by

$$E_1 : \{K_1 > (M + N - r_1)r_1\}. \tag{5.33}$$

Similarly, $E_2$ denotes the event in which the matrix $\mathbf{M}_2$ is recovered, is given by

$$E_2 : \{K_2 > (M + N - r_2)r_2\}. \tag{5.34}$$

Finally, denote by $E_3$ the event in which the matrix $\mathbf{M}$ is recovered

$$E_3 : \{K_1 + K_2 > (2M + N - r)r\}. \tag{5.35}$$

**Theorem 3.** *Let* $\mathbf{M}_1 \in \mathbb{R}^{M\times N}$ *and* $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$*, with* $\operatorname{rank}(\mathbf{M}_1) = r_1$ *and* $\operatorname{rank}(\mathbf{M}_2) = r_2$*. When the locations of the missing entries are sampled uniformly at random with*

*probability $p_1$ and $p_2$ respectively, the probability of recovery is bounded by*

$$\Pr[E_i] \geq \begin{cases} 1 - e^{-2\frac{(MNp_i - k_i)^2}{MN}}, & \text{for } p_i > \frac{k_i}{MN} \\ 0, & \text{for } p_i \leq \frac{k_i}{MN}, \end{cases} \tag{5.36}$$

*where $E_i$ is the event in which the matrix $\mathbf{M}_i$ is recovered and $i \in \{1, 2\}$.*

*Proof.* See Appendix A. □

In view of this, the probability of recovering both matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ independently is given by

$$\Pr[E_1 \wedge E_2] = \begin{cases} (1 - e^{-2\frac{(MNp_1 - k_1)^2}{MN}})(1 - e^{-2\frac{(MNp_2 - k_2)^2}{MN}}), & \text{for } p_i > \frac{k_i}{MN}, i \in \{1, 2\} \\ 0, & \text{otherwise.} \end{cases} \tag{5.37}$$

**Theorem 4.** *Let $\mathbf{M}_1 \in \mathbb{R}^{M \times N}$ and $\mathbf{M}_2 \in \mathbb{R}^{M \times N}$, with $\operatorname{rank}(\mathbf{M}_1) = r_1$ and $\operatorname{rank}(\mathbf{M}_2) = r_2$. Given the combined matrix $\mathbf{M} = [\mathbf{M}_1^T \, \mathbf{M}_2^T]^T$ of rank $r$, when the missing entries from $\mathbf{M}_1$ and $\mathbf{M}_2$ are sampled uniformly at random with probability $p_1$ and $p_2$, respectively, the probability of recovering the matrix $\mathbf{M}$ is bounded by*

$$\Pr[E_3] \geq \begin{cases} 1 - e^{-\frac{(MN(p_1 + p_2) - k)^2}{MN}}, & \text{for } p_1 + p_2 > \frac{k}{MN} \\ 0, & \text{for } p_1 + p_2 \leq \frac{k}{MN}, \end{cases} \tag{5.38}$$

*where $E_3$ is the event in which the matrix $\mathbf{M}$ is recovered and is defined in (5.35).*

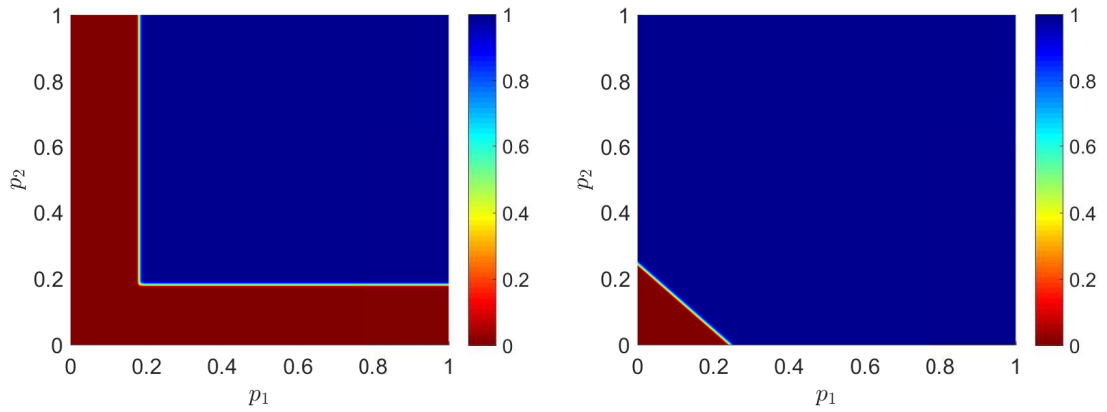
*Proof.* See Appendix A. □

A numerical comparison of the two recovery probabilities follows.

### 5.4.1 Numerical evaluation of the recovery probabilities

This section provides a numerical comparison for the independent and joint recovery probabilities for different rank values. In particular, the size of the matrices is fixed to $M = 50$ and $N = 100$ and the rank values considered are depicted in Table 5.1.

Figure 5.5a shows the probability of recovering both datasets independently depending on the sampling probabilities of the two matrices, when all three rank values are equal to 6. Note that the sampling is uniformly at random with probability $p_1$ for the entries $\mathbf{M}_1$ and probability $p_2$ for the entries in $\mathbf{M}_2$. Interestingly, in the probabilistic framework there is a smooth transition between the region with probability one and the region in which the recovery is not possible. It is also worth




(a) Independent recovery probability (b) Joint recovery probability

Figure 5.5. Recovery probability as a function of $p_1$ and $p_2$ when $r_1 = 6$, $r_2 = 6$ and $r = 6$.

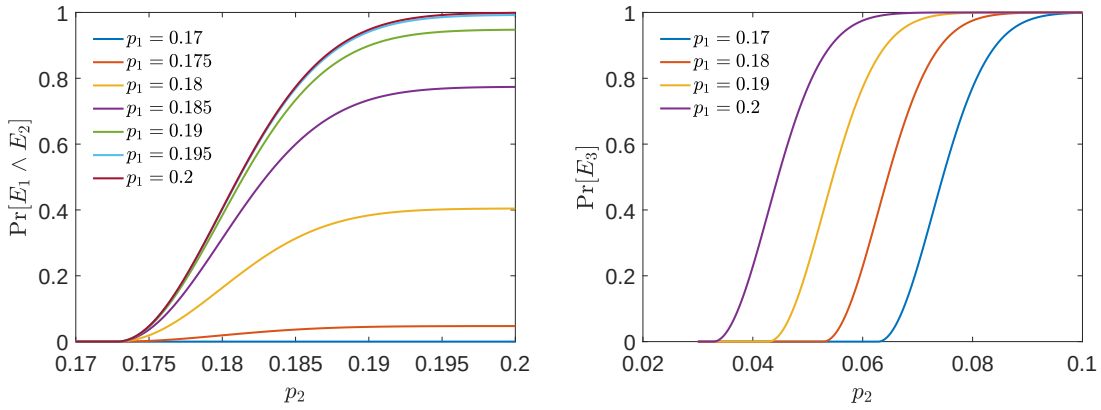
mentioning that the transition between regions is fast. The probability of recovery increases rapidly to one as the sampling probabilities increase.

The probability of the joint recovery is depicted in Figure 5.5b when all three rank values are equal to 6. Interestingly, the shape of the region with probability one is the same as in Figure 5.1. It is also worth noting that the transition between the no recovery region and the probability one region is fast.

Comparing Figures 5.5a and 5.5b, it is interesting to note that the number of cases in which the joint recovery is possible is much larger compared to the independent recovery case. In particular, when $r_1 = 6$, $r_2 = 6$ and $r = 6$ the area with probability one is smaller in the independent case compared to the joint recovery case. Using Definition 7, the existence of beneficial joint recovery depends on the existence of the region in which the joint recovery requires fewer entries compared to the independent case.

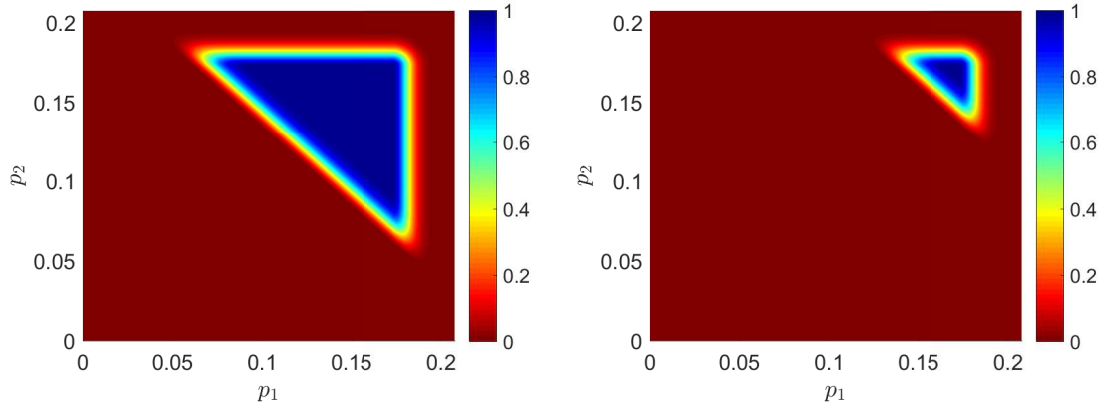
Figure 5.6a shows the transition between the region with probability zero and the region with probability one for the independent recovery event when $r_1 = 6$, $r_2 = 6$ and $r = 6$. Interestingly, when $p_1 \leq 0.19$ the probability of the independent recovery never converges to one regardless of the number of observations from the second dataset. In is also worth noting that the probability increases exponential from zero to one for $p_1 > 0.19$. In fact, the width of the transition region is 0.02 in sampling probability for both datasets.

Figure 5.6b depicts the transition between the region with probability zero and the region with probability one for the joint recovery event when $r_1 = 6$, $r_2 = 6$ and $r = 6$. In contrast with the independent recovery case, the probability always converges to one as one of the sampling probability increases. It is worth noting that the increase in recovery probability is exponential and the width of the transition region is approximately 0.02 in sampling probability.




(a) Independent recovery probability (b) Joint recovery probability

Figure 5.6. Recovery probability bound as a function of $p_2$ when $p_1$ is fixed and $r_1 = 6$, $r_2 = 6$ and $r = 6$.

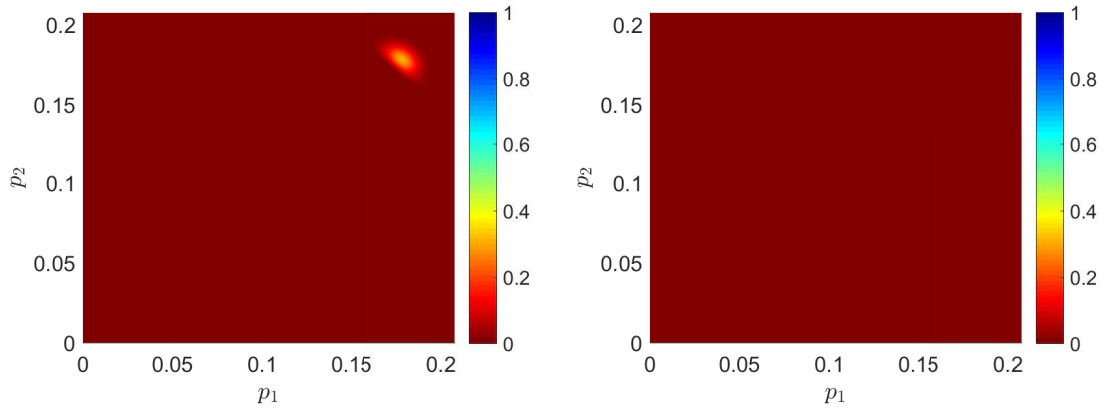



(a) $r_1 = 6$, $r_2 = 6$ and $r = 6$ (b) $r_1 = 6$, $r_2 = 6$ and $r = 8$

Figure 5.7. Probability of the event in which the joint recovery is beneficial as a function of $p_1$ and $p_2$ for different values of rank.

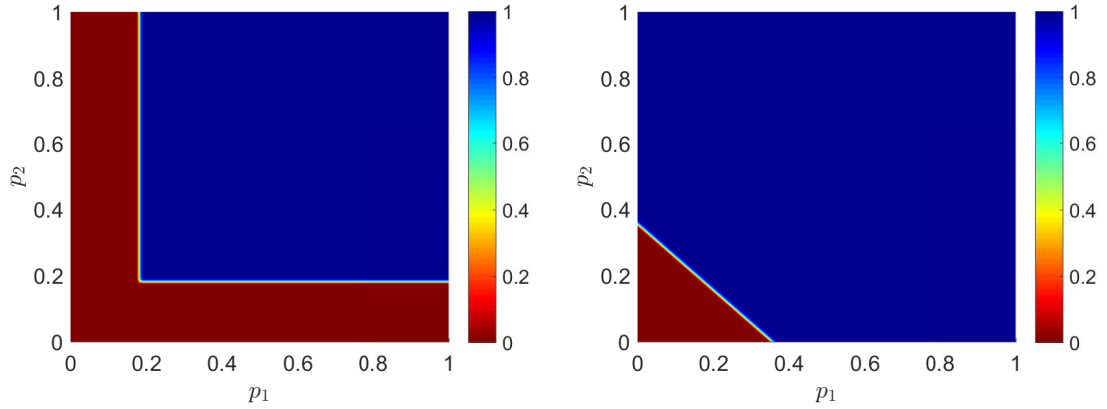
Figure 5.7a depicts the probability of the event in which the joint recovery requires fewer entries compared to the independent recovery. Interestingly, the shape of the region with probability one is similar to $\mathcal{R}_7$. The same triangle shape is observed in the probabilistic framework. Also, the transition between the 0 and the 1 probability areas is fast.

As expected, when the rank of the combined matrix increases, the size of the region in which the joint recovery requires fewer entries compared to the independent recovery decreases. For instance, Figure 5.8a depicts the probability of the event in which the joint recovery is beneficial when $r_1 = 6$, $r_2 = 6$ and $r = 9$. In comparison to Figure 5.7a, the increase of $r$ from 6 to 9 vanished the region with probability one. However, since the value $r = 9$ is smaller than the threshold in Theorem 2 the probability of the event in which the joint recovery is beneficial is not always zero.




(a) $r_1 = 6$, $r_2 = 6$ and $r = 9$ (b) $r_1 = 6$, $r_2 = 6$ and $r = 10$

Figure 5.8. Probability of the event in which the joint recovery is beneficial as a function of $p_1$ and $p_2$ for different values of rank.



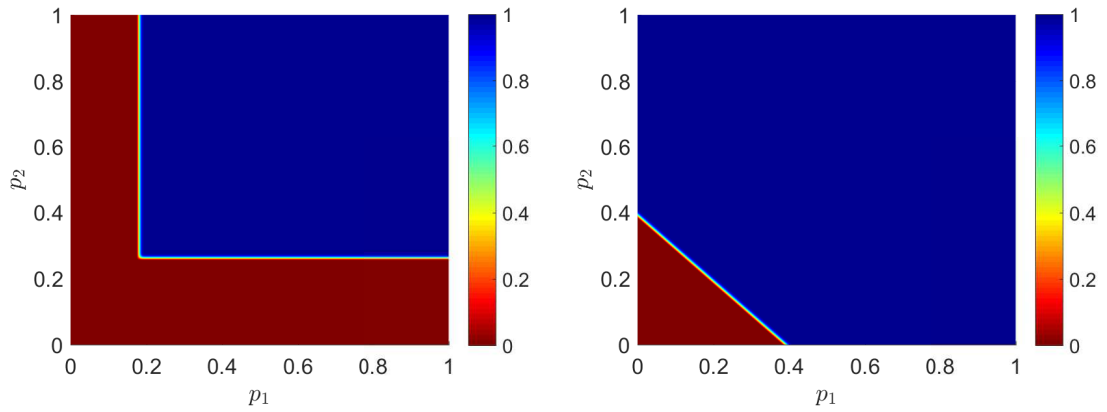

(a) Independent recovery probability (b) Joint recovery probability

Figure 5.9. Recovery probability as a function of $p_1$ and $p_2$ when $r_1 = 6$, $r_2 = 6$ and $r = 9$.

However, for $r_1 = 6$, $r_2 = 6$ and values of $r$ greater than the threshold in Theorem 2, i.e., $r > 9.05$, the probability of the event in which the joint recovery is beneficial is zero. Figure 5.8b shows the probability of the event in which the joint recovery is beneficial for $r = 10$. As expected, when the rank of the condition in Theorem 2 is not satisfied, the joint recovery is not advantageous.

Figure 5.9a depicts the probability of independent recovery of the two matrices when $r_1 = 6$, $r_2 = 6$ and $r = 9$. Not surprisingly, the probability is the same as in Figure 5.5a since the independent recovery depends only on the individual rank of the two matrices. However, the probability of the joint recovery is different since the rank of the combined matrix is different.

Figure 5.9b shows the probability of the joint recovery event as a function of the sampling probabilities for the two datasets. In comparison to Figure 5.5b, the area




(a) Independent recovery probability (b) Joint recovery probability

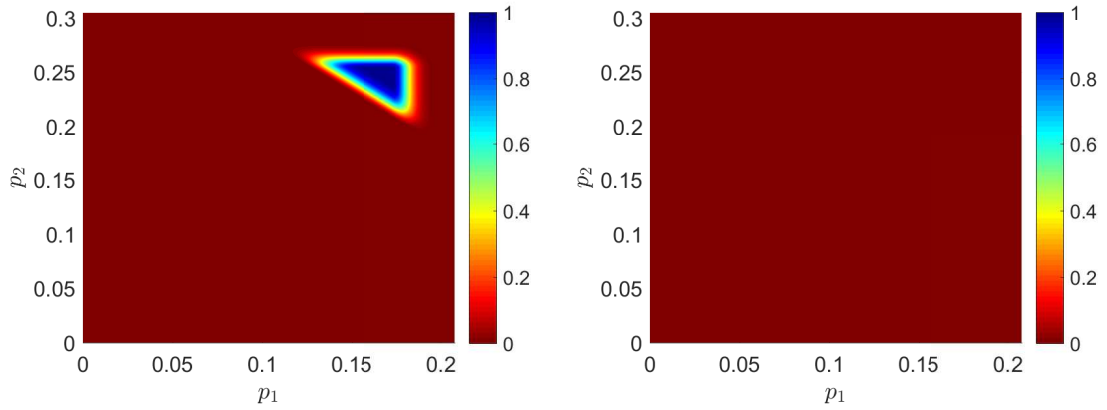
Figure 5.10. Recovery probability as a function of $p_1$ and $p_2$ when $r_1 = 6$, $r_2 = 9$ and $r = 10$.

with probability one is smaller. As expected, when the rank of the combined matrix increases, the number of cases in which the joint recovery is possible decreases.

In addition to the rank of the combined matrix, the probability of the event in which the joint recovery is beneficial depends on the ratio between the rank of the two matrices combined. This observation also is mentioned in Section 5.3.2. To assess the impact of different rank ratios in the probabilistic framework, different values of $r_1$ and $r_2$ are considered.

Figure 5.10a depicts the probability of the independent recovery of $\mathbf{M}_1$ and $\mathbf{M}_2$ when $r_1 = 6$, $r_2 = 9$ and $r = 10$. As expected, the region with probability one is similar to $\mathcal{R}_4$ in Figure 5.4. In comparison to Figure 5.5a where the region with probability one is square, in Figure 5.10a the region is rectangular. In fact, the shape of region $\mathcal{R}_4$ is given by the ratio between $\underline{k}_1$ and $\underline{k}_2$ which ultimately depends on the ratio between $r_2$ and $r_1$. However, the transition between the zero probability and the probability one region is still fast.

Figure 5.10b shows the probability of the joint recovery when $r_1 = 6$, $r_2 = 9$ and $r = 10$. As expected, the probability of joint recovery is similar to Figure 5.9b because it only depends on the value of $r$. Regardless of the values of $r_1$ and $r_2$ for the same value of $r$ the probability of joint recovery is the same. However, the probability of the event in which the joint recovery is beneficial depends on all three rank values. To showcase this, Figure 5.11a and Figure 5.8b depict the probability of the event in which the joint recovery is beneficial for the same value of rank for the combined matrix but different rank values for $\mathbf{M}_1$ and $\mathbf{M}_2$. In contrast to the other case, when $r_1 = 6$ and $r_2 = 9$ the conditions of Theorem 2 are satisfied and the joint recovery is beneficial. Moreover, the same triangle shape is observed for the region with probability one for the event in which the joint recovery is beneficial. Interestingly, the transition between the probability one region and the region with




(a) $r_1 = 6$, $r_2 = 9$ and $r = 10$ (b) $r_1 = 6$, $r_2 = 9$ and $r = 12$

Figure 5.11. Probability of the event in which the joint recovery is beneficial as a function of $p_1$ and $p_2$ for different values of rank.

probability zero is faster along the hypotenuse and slower along the catheti. This suggests that when sampling probabilities are increased to boost the probability of recovery it is more efficient to increase both sampling probabilities rather than having a more substantial increase in only one of them.

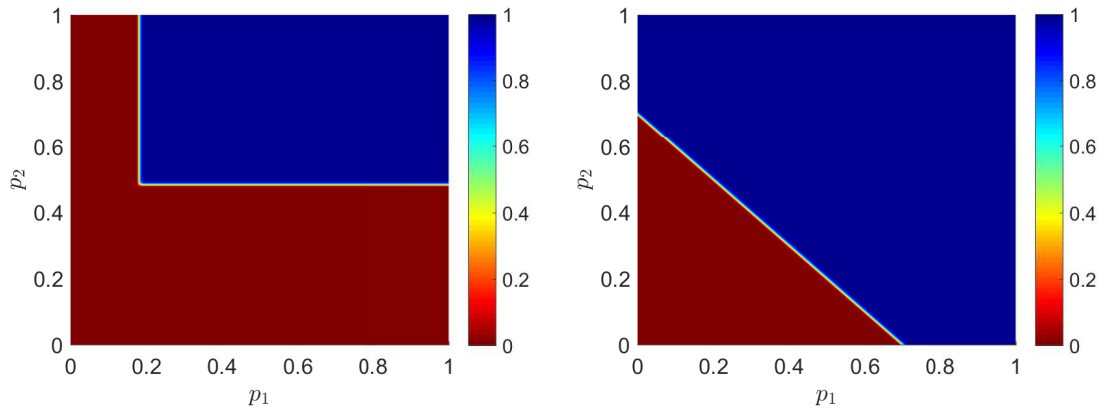
In addition, similar to the case in which $r_1$ and $r_2$ are equal, the size of the triangle decreases as the rank of the combined matrix increases. Figure 5.11b shows the probability of the event in which the joint recovery is beneficial when $r_1 = 6$, $r_2 = 9$ and $r = 12$. Note that the probability converges to zero when the rank of the combined matrix becomes larger that the threshold in Theorem 2.

Since the probability of the event in which the joint recovery is beneficial also depends on the ratio between $r_2$ and $r_1$, it is expected that for a large enough ratio the probability will converge to zero. In this context, the following set of results consider $r_1 = 6$ and $r_2 = 18$.

Figure 5.12a depicts the probability of the independent recovery when $r_1 = 6$, $r_2 = 18$ and $r = 19$. As expected, the shape of the region with probability one is rectangular. Also the area is smaller compared to the previous cases.

For the joint recovery, Figure 5.12b shows the probability for the same rank values. It is worth mentioning that even in this case the area with probability one is larger in the joint recovery case compared to the independent recovery case. However, as suggested by the results in Table 5.1, the joint recovery is not beneficial. This is because the rank upper bound in Theorem 2 is larger than the lower bound for $r$.

Interestingly, the figures with the probability of recovery present similar characteristics with the corresponding diagrams presented in Section 5.3.2. In addition, the numerical results are consistent with the sufficient condition for the joint recovery to be beneficial in Theorem 2. Specifically, when the rank of the combined matrix is larger than the threshold, the probability of the joint recovery event converges




(a) Independent recovery probability (b) Joint recovery probability

Figure 5.12. Recovery probability as a function of $p_1$ and $p_2$ when $r_1 = 6$, $r_2 = 18$ and $r = 19$.

to zero (see Figure 5.8b). The next section exploits the probabilistic bounds in an analytical framework for optimizing the sampling process.

## 5.5 Sample complexity minimization for joint recovery of multiple datasets

This section analyses the problem of minimizing the cost of recovering missing data from two datasets. Consider, two datasets described by the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ respectively. In general, the cost induced by the acquisition process is different for each monitoring system. Typically, the acquisition cost depends on the type of the sensors used, the cost of maintaining the sensors, etc. Therefore, different types of data have different acquisition costs.

Coming back to the example presented in the introduction, traffic flow data can be collected using the Motorway Incident Detection and Automatic Signalling (MIDAS) system which consists of induction loops spaced at 500 meter intervals in the road [114]. Moreover, the cost for installing MIDAS are in the order of £210,000 per km [114]. Assuming that there are three lanes on each way, for 1 km of road there are 12 induction loops. Consequently, the average cost per sensor is of £17,500. On the other hand, the advances in air quality monitoring tools lead to a cost of £2,000 per sensor [115]. In this context, the ratio between the cost of a $CO_2$ level observation and the traffic flow observation is 0.114. In other words, the acquisition cost for traffic flow data is 8.75 times higher than that of air quality data. Given the significant difference between the two sampling costs, it makes sense to optimize the total sampling cost for the recovery of both types of data.

In the two datasets case under consideration, denote by $\alpha_1$ the cost of acquiring one observation from the first dataset and by $\alpha_2$ the cost of acquiring one observation from the second dataset. Since not all the entries in $\mathbf{M}_1$ and $\mathbf{M}_2$ are available, the expected cost of sampling the two matrices, denoted by $\beta$, and is given by

$$\beta = \alpha_1 p_1 + \alpha_2 p_2, \tag{5.39}$$

where $p_1$ is the probability of sampling the entries from $\mathbf{M}_1$ and $p_2$ is the probability of sampling the entries from $\mathbf{M}_2$. Note that the cost $\beta$ depends on the acquisition costs for both datasets but also on the sampling probabilities. In this section, the sampling of the two datasets is assumed to be uniformly distributed. This facilitates the use of the bounds derived for the probabilities of independent (5.37) and joint (5.38) recovery in Section 5.4.

The minimum sampling cost depends on the probability of recovery. The objective is to determine the minimum cost $\beta$ that allows the recovery of both matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ with a given probability of recovery $q$. Note that the two matrices can be recovered independently or jointly. In this context, the optimization parameters are the two sampling probabilities $p_1$ and $p_2$. The solution of the optimization problem depends on the sampling costs of each dataset and on the recovery probability.

### 5.5.1 Independent recovery

The problem of minimizing the cost of recovering both matrices independently is formulated as

$$\begin{aligned} \underset{p_1,p_2}{\text{minimize}} \quad & \alpha_1 p_1 + \alpha_2 p_2 \\ \text{subject to} \quad & \Pr[E_1] \geq q_1 \\ & \Pr[E_2] \geq q_2, \end{aligned} \tag{5.40}$$

where $E_1$ and $E_2$ are the events in which the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are recovered as defined in (5.33) and (5.34), respectively. Moreover, the probabilities of the events $E_1$ and $E_2$ are lower bounded in (A.137). The following lemma uses the result in (A.137) to bound the sampling probabilities based on the probability of recovery.

**Lemma 3.** *Matrix* $\mathbf{M}_i \in \mathbb{R}^{M\times N}$ *with* $\operatorname{rank}(\mathbf{M}_i) = r_i$ *is recovered with probability at least* $q_i$ *if the sampling is uniformly at random with probability* $p_i$ *bounded by*

$$p_i \geq \frac{2MN(M+N-r_i)r_i + \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2}, \tag{5.41}$$

*for* $i \in \{1,2\}$.

*Proof.* See Appendix A. □

In other words, Lemma 3 bounds the sampling probability based on the probability of recovering the matrix. Using Lemma 3, the optimization problem in (5.40) can be recast as

$$\begin{aligned} \underset{p_1,p_2}{\text{minimize}} \quad & \alpha_1 p_1 + \alpha_2 p_2 \\ \text{subject to} \quad & p_1 \geq \frac{2MN(M+N-r_1)r_1 + \sqrt{-2M^3N^3\ln(1-q_1)}}{2M^2N^2} \\ & p_2 \geq \frac{2MN(M+N-r_2)r_2 + \sqrt{-2M^3N^3\ln(1-q_2)}}{2M^2N^2}. \end{aligned} \tag{5.42}$$

Note that because the samping costs $\alpha_1$ and $\alpha_2$ are positive, $\beta$ is a monotonically increasing function on $p_1$ and $p_2$. Therefore minimizing the sampling cost $\beta$ is equivalent to minimizing the sampling probabilities $p_1$ and $p_2$. Based on this observation, the problem in (5.42) is equivalent to

$$\begin{aligned} \text{minimize} \quad & p_1, p_2 \\ \text{subject to} \quad & p_1 \geq \frac{2MN(M+N-r_1)r_1 + \sqrt{-2M^3N^3\ln(1-q_1)}}{2M^2N^2} \\ & p_2 \geq \frac{2MN(M+N-r_2)r_2 + \sqrt{-2M^3N^3\ln(1-q_2)}}{2M^2N^2}. \end{aligned} \tag{5.43}$$

Interestingly, when $q_1 \in (0,1)$ and $q_2 \in (0,1)$ the lower bounds for $p_1$ and $p_2$ are monotonically increasing functions in $q_1$ and $q_2$ respectively. In this case, the solution to the problem in (5.43) is given by

$$\begin{cases} p_1 = \frac{2MN(M+N-r_1)r_1+\sqrt{-2M^3N^3\ln(1-q_1)}}{2M^2N^2}, \\ p_2 = \frac{2MN(M+N-r_2)r_2+\sqrt{-2M^3N^3\ln(1-q_2)}}{2M^2N^2}. \end{cases} \tag{5.44}$$

Consequently, this is also the solution to the initial problem in (5.40). Denote by $\beta_i$ the minimum sampling cost for the independent recovery of the matrix $\mathbf{M}_1$ with probability at least $q_1$ and the matrix $\mathbf{M}_1$ with probability at least $q_2$. The index $i$ for the cost $\beta_i$ refers to the independent recovery case. Using the solution obtained in (5.44), the minimum cost for the independent recovery case is given by

$$\begin{aligned} \beta_i = & \alpha_1 \frac{2MN(M+N-r_1)r_1 + \sqrt{-2M^3N^3\ln(1-q_1)}}{2M^2N^2} + \\ & \alpha_2 \frac{2MN(M+N-r_2)r_2 + \sqrt{-2M^3N^3\ln(1-q_2)}}{2M^2N^2}. \end{aligned} \tag{5.45}$$

As expected, the minimum sampling cost for the independent recovery depends on the sampling costs $\alpha_1$ and $\alpha_2$ and the minimum sampling probabilities that guarantee

the recovery of $\mathbf{M}_1$ with probability at least $q_1$ and $\mathbf{M}_2$ with probability at least $q_2$. To assess the benefit of joint recovery the next section analyses the minimum sampling cost for recovering the matrix $\mathbf{M}$.

### 5.5.2 Joint recovery

Using the formulation in (5.40) the problem of minimizing the sampling cost of recovering both matrices jointly is defined as

$$\begin{aligned} \underset{p_1,p_2}{\text{minimize}} \quad & \alpha_1 p_1 + \alpha_2 p_2 \\ \text{subject to} \quad & \Pr[E_3] \geq q, \end{aligned} \tag{5.46}$$

where $E_3$ is defined in (5.35) and denotes the event in which the matrix $\mathbf{M} = \left[\mathbf{M}_1^T \, \mathbf{M}_2^T\right]^T$ is recovered. In addition, the probability of the event $E_3$ is bounded based on the sampling probabilities of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ in (5.38). The following lemma bounds the sampling probabilities such that the matrix $\mathbf{M}$ is recovered with probability at least $q$.

**Lemma 4.** *Matrix* $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2M \times N}$ *with* $\operatorname{rank}(\mathbf{M}) = r$ *is recovered with probability at least* $q$ *if the matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are sampled uniformly at random with probabilities* $p_1$ *and* $p_2$*, respectively, and the sampling probabilities satisfy*

$$p_1 + p_2 \geq \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3 \ln(1-q)}}{2M^2N^2}. \tag{5.47}$$

*Proof.* See Appendix A. □

Lemma 4 bounds the sampling probabilities for $\mathbf{M}_1$ and $\mathbf{M}_2$ based on the probability of recovering the matrices jointly. Using Lemma 4, the optimization problem in (5.46) becomes

$$\begin{aligned} \underset{p_1,p_2}{\text{minimize}} \quad & \alpha_1 p_1 + \alpha_2 p_2 \\ \text{subject to} \quad & p_1 + p_2 \geq \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3 \ln(1-q)}}{2M^2N^2}. \end{aligned} \tag{5.48}$$

Note that when the sampling costs $\alpha_1$ and $\alpha_2$ are positive, the total cost is a monotonically increasing function on $p_1$ and $p_2$. Therefore, minimizing the sampling cost is equivalent to minimizing the sampling probabilities $p_1$ and $p_2$ subject to the constraints imposed by the fundamental limit. Thus, the optimization problem in

(5.48) is equivalent to

$$\begin{aligned} &\text{minimize} \quad p_1, p_2 \\ &\text{subject to} \quad p_1 + p_2 \geq \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \end{aligned} \tag{5.49}$$

Moreover, when $q \in (0,1)$, the lower bound for $p_1 + p_2$ is a monotonically increasing function on $q$, and therefore, the solution to the optimization problem in (5.49) satisfies

$$p_1 + p_2 = \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \tag{5.50}$$

In contrast to the independent recovery case, for the joint recovery case there is an infinite number of values of $p_1$ and $p_2$ that satisfy the constraint in (5.50). However, the total sampling cost for the joint recovery case is not minimized for all values of $p_1$ and $p_2$ that satisfy (5.50). Note that the cost $\beta$ is also a function of the independent sampling costs of the two datasets. Therefore, in the cost minimization problem presented in (5.46) and (5.49), the choice of $p_1$ and $p_2$ that satisfy (5.50) is governed by the values of $\alpha_1$ and $\alpha_2$.

Denote by $\beta_j$ the minimum sampling cost achievable when recovering the missing entries in a joint setting. In other words, $\beta_j$ is the solution to the optimization problems in (5.46) and (5.49). Therefore, $\beta_j$ is given by

$$\beta_j = \begin{cases} \alpha_1 \frac{2MN(M+N-r)r+\sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}, & \text{for } \alpha_1 \leq \alpha_2 \\ \alpha_2 \frac{2MN(M+N-r)r+\sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}, & \text{for } \alpha_1 > \alpha_2. \end{cases} \tag{5.51}$$

The minimum cost is achieved by sampling only the dataset with the smaller sampling cost. For the independent recovery, the pair ($p_1$,$p_2$) that satisfies the conditions to recover the two matrices is unique. In contrast, for the joint recovery there exists an infinite number of pairs ($p_1$,$p_2$) that satisfy the condition on the probability of recovering the matrix $\mathbf{M}$. The flexibility induced by the joint recovery framework facilitates the use of asymmetric sampling strategies for the datasets combined. A numerical comparison of the minimum sampling costs induced by the independent and joint recovery follows.

### 5.5.3 Numerical comparison for the minimum sampling cost

This section provides a numerical comparison for the minimum sampling cost achievable in the independent and joint recovery cases. In particular, the optimization problems in (5.40) and (5.46) for the independent and joint recovery respectively, are solved for the rank values depicted in Table 5.1 when $M = 50$ and $N = 100$.

For convenience, the solutions for the two optimization problems are recalled in this section. The minimum sampling cost for the independent recovery, denoted by $\beta_i$, is given by

$$\begin{aligned}\beta_i =&\alpha_1 \frac{2MN(M+N-r_1)r_1 + \sqrt{-2M^3N^3\ln(1-q_1)}}{2M^2N^2} + \\ &\alpha_2 \frac{2MN(M+N-r_2)r_2 + \sqrt{-2M^3N^3\ln(1-q_2)}}{2M^2N^2},\end{aligned} \tag{5.52}$$

where $q_i$ is the minimum probability of recovering the matrix $\mathbf{M}_i$, with $i \in \{1, 2\}$. Also, the minimum sampling cost for the joint recovery, denoted by $\beta_j$, is

$$\beta_j = \begin{cases} \alpha_1 \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}, & \text{for } \alpha_1 \leq \alpha_2 \\ \alpha_2 \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}, & \text{for } \alpha_1 > \alpha_2, \end{cases} \tag{5.53}$$

where $q$ is the minimum probability of recovering the matrix $\mathbf{M}$.

Note that dividing the expected cost in equation (5.39) by $\alpha_1$ yields

$$\frac{1}{\alpha_1}\beta = p_1 + \frac{\alpha_2}{\alpha_1}p_2. \tag{5.54}$$

Given that the sampling cost $\alpha_1$ is assumed to be constant for a given dataset, minimizing the expected cost is equivalent to minimizing the term $\frac{1}{\alpha_1}\beta$. Therefore, the minimum expected cost depends on the ratio between the two sampling costs. A numerical evaluation of both the independent and the joint recovery costs is presented below for different ratios of $\alpha_1$ and $\alpha_2$.

Figure 5.13 depicts a numerical comparison between the minimum sampling costs for the independent and joint recovery, when $r_1 = 6$ and $r_2 = 6$, and for different values of $r$. Interestingly, when the difference between the two sampling costs is significant, the minimum cost for the joint recovery is strictly smaller than the minimum cost for the independent recovery. In contrast, when the ratio between the two sampling costs is close to one, the difference between $\beta_i$ and $\beta_j$ is smaller. It is also worth mentioning that when the value of $r$ is smaller than the threshold given by Theorem 2, the joint recovery costs less than the independent recovery of the two matrices. On the other hand, for values of $r$ greater that the limit given by Theorem 2, there are cases in which recovering the two matrices independently is cheaper. In addition, the number of cases in which the joint recovery costs more increases as the rank of the combined matrix increases. In other words, the less correlated the two datasets are the less advantageous is to recover them jointly.

Figure 5.14 compares for the minimum sampling costs when $r_1 = 6$ and $r_2 = 9$ and $r$ varies according to Table 5.1. Similar to the case in which $r_1 = 6$ and $r_2 = 6$,

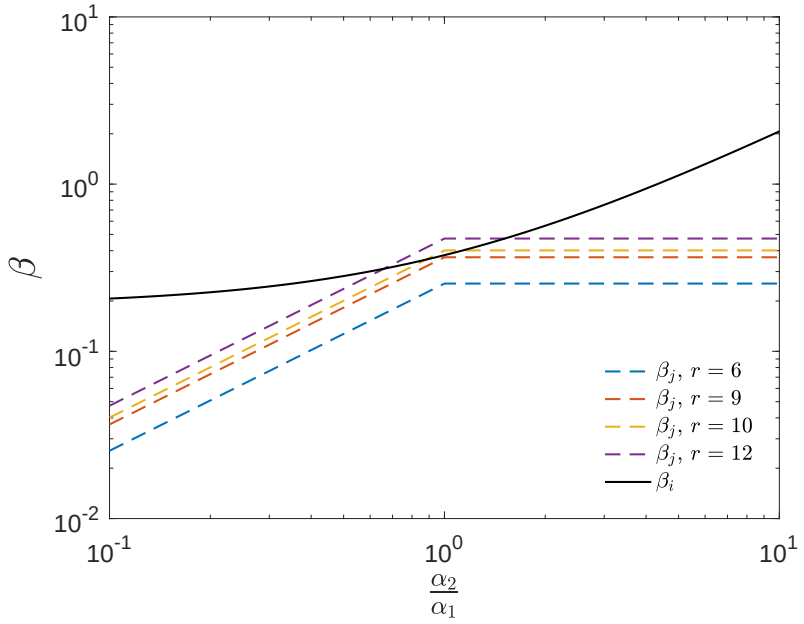



Figure 5.13. Numerical comparison of the minimum sampling cost for the independent and joint recovery depending on the ratio between the sampling costs, when $r_1 = 6$ and $r_2 = 6$, and for different values of $r$.

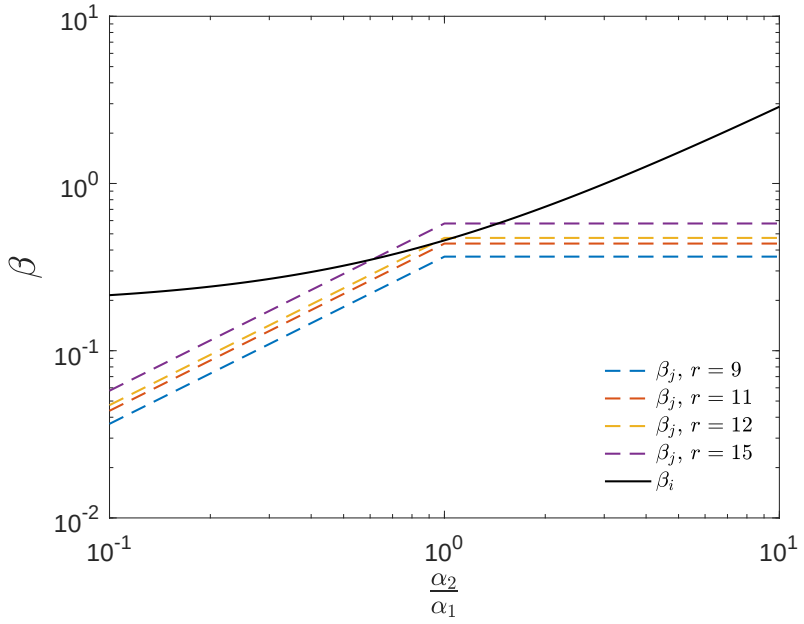



Figure 5.14. Numerical comparison of the minimum sampling cost for the independent and joint recovery depending on the ratio between the sampling costs, when $r_1 = 6$ and $r_2 = 9$, and for different values of $r$.

the joint recovery costs less when the difference between the two sampling costs is larger. However, when the ratio between the two sampling costs is close to one the cheapest option depends on the value of $r$ . Precisely, when $r$ is smaller than the

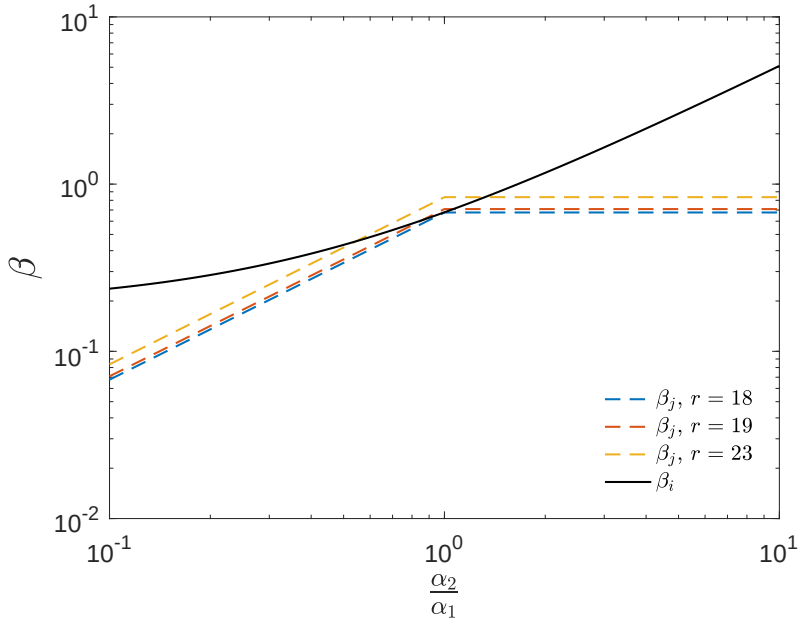



Figure 5.15. Numerical comparison of the minimum sampling cost for the independent and joint recovery depending on the ratio between the sampling costs, when $r_1 = 6$ and $r_2 = 18$, and for different values of $r$.

threshold in Theorem 2, the joint recovery always costs less. On the other hand, the independent recovery is advantageous when $r$ is larger than the limit in Theorem 2 and the ratio between the two sampling costs is close to one. Based on the numerical results depicted in Figures 5.13 and 5.14, the joint recovery costs less than the independent recovery when the conditions for the joint recovery to be beneficial are satisfied.

Figure 5.15 shows the minimum sampling cost for the independent and joint recovery when $r_1 = 6$ and $r_2 = 18$, and for different values of $r$. Note that for this case, the joint recovery is not beneficial as the upper bound given by Theorem 2 is smaller than the lower possible value for $r$. Therefore, the minimum cost of the joint recovery is not always smaller than the minimum cost for the independent recovery. However, consistent with the other results in this section, the joint recovery induces a smaller sampling cost when the difference between the two sampling costs is large.

In the example presented in the introduction, the ratio between the acquisition cost for the $CO_2$ level data and the traffic flow data is $\alpha_{12} = 0.114$. In view of this, regardless of the ratio between the rank of the two matrices, the joint recovery of the two datasets induces a smaller total sampling cost.

The numerical results presented in this section address the problem of minimizing the sampling cost for the recovery of two datasets. In particular, the minimum sampling cost achievable in the independent recovery is compared with the minimum sampling cost for the joint recovery. Based on the numerical results presented, the

joint recovery costs less when the difference between the two sampling costs is large. In contrast, when the ratio between the two sampling costs is close to one, the joint recovery induces a smaller sampling costs when the conditions in Theorem 2 are satisfied.

## 5.6 Numerical analysis

This section presents a numerical evaluation of the joint recovery performance for two datasets. The size of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ is fixed such that $M = 50$ and $N = 100$. Consequently, the joint matrix $\mathbf{M}$ is a square matrix of size 100. In this context, different values of rank are assessed.

Table 5.1 contains the values of rank under consideration throughout the section. The range of rank values selected aims to characterize the joint recovery in three scenarios:

- when the two combined matrices have the same rank

$$\frac{r_2}{r_1} = 1, \tag{5.55}$$

- when the ratio between the two rank values is small

$$1 < \frac{r_2}{r_1} \leq 2, \tag{5.56}$$

- when the ratio between the two rank values is large

$$\frac{r_2}{r_1} > 2. \tag{5.57}$$

Consequently, the ratios between the rank of the combined matrices considered are $r_2/r_1 = 1$, $r_2/r_1 = 1.5$ and $r_2/r_1 = 3$. As noted in Section 5.3.2, the ratio the two rank values is one of the important factors for guaranteeing that the joint recovery is beneficial. When the ratio between the rank of the first and the second matrix is large, the rank condition in (5.31) from Theorem 2 is not satisfied. Indeed, Table 5.1 shows that when $r_1 = 6$ and $r_2 = 18$, the upper bound given by Theorem 2, i.e., 17.78, is smaller than the lower bound for the rank of the combined matrix, i.e., 18. In this case, the joint recovery is not beneficial due to the large ratio between the two rank values. In contrast, when the two rank values are equal, i.e., $r_1 = 6$ and $r_2 = 6$, the rank condition in Theorem 2 is satisfied in most of the rank possibilities for $r$. Notably, the joint recovery is beneficial in four cases out of the total of seven possible values for $r$. As the ratio between $r_2$ and $r_1$ increases, the limit in Theorem 2 shifts towards the lower bound for the rank of the combined matrices. Therefore,

the number of cases in which the joint recovery is beneficial decreases. In particular, when $r_1 = 6$ and $r_2 = 9$, out of the seven possible values for the rank for the combined matrix, the joint recovery is beneficial in only three cases. Note the decrease in the number of cases compared to the scenario in which $r_1 = r_2$.

The numerical results in this section are obtained by simulating the recovery with synthetic data. The following section describes the model used to generate the datasets used for the recovery performance comparison. The choice for synthetic data rather that real data is mainly motivated by the flexibility of having datasets with the rank values described in Table 5.1. Therefore, the use of synthetic generated datasets facilitates the comparison between SVT and BSVT-based recovery for all rank values in Table 5.1. However, not all the numerical results are included in this section. To avoid redundancy, only numerical results that provide additional insight into the joint recovery performance are included.

### 5.6.1 Simulation framework

A mathematical description of the model used to generate synthetic correlated datasets follows. The combined state variable matrix is defined as

$$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_N], \tag{5.58}$$

where each state variable vector $\mathbf{m}_i \in \mathbb{R}^{2M}$ for $i \in \{1, 2, ..., N\}$ is generated by a multivariate Gaussian process with $\mathbf{0}$ mean and covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{m}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}). \tag{5.59}$$

The covariance matrix $\mathbf{\Sigma}$ is a block matrix in which block $\mathbf{\Sigma}_{ii}$ is a Toeplitz matrix describing the covariance matrix of the dataset $i$. The resulting covariance matrix is given by

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}, \tag{5.60}$$

where $\mathbf{\Sigma}_{ml} \in \mathbb{R}^{M \times M}$ for $m, l \in \{1, 2\}$. In this framework, the elements of $\mathbf{\Sigma}_{ml}$ are defined as

$$(\mathbf{\Sigma}_{ml})_{i,j} = \rho^{\frac{1}{\zeta_{ml}}|i-j|}, \tag{5.61}$$

where $\rho \in (0, 1)$. Therefore, the values on the diagonals of the Toeplitz matrix $\mathbf{\Sigma}_{ml}$ are exponentially decreasing from 1 to $\upsilon_{ml}$ where

$$\upsilon_{ml} = \rho^{\frac{1}{\zeta_{ml}}(M-1)}. \tag{5.62}$$

Precisely, the matrix $\mathbf{\Sigma}_{ml}$ is given by

$$\mathbf{\Sigma}_{ml} = \begin{bmatrix} 1 & \rho^{\frac{1}{\zeta_{ml}}} & \cdots & \rho^{\frac{1}{\zeta_{ml}}(M-1)} \\ \rho^{\frac{1}{\zeta_{ml}}} & 1 & \cdots & \rho^{\frac{1}{\zeta_{ml}}(M-2)} \\ & & \ddots & \\ \rho^{\frac{1}{\zeta_{ml}}(M-1)} & \rho^{\frac{1}{\zeta_{ml}}(M-2)} & \cdots & 1 \end{bmatrix}. \tag{5.63}$$

The parameter $\upsilon_{ml}$ uniquely determines the value of $\zeta_{ml}$. In order to guarantee that the covariance matrix is symmetric and positive semi-definite, the following constrains are applied to the model

$$\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21} = \psi\mathbf{\Sigma}_{11}, \tag{5.64}$$

where $0 \leq \psi \leq 1$. Consequently, the covariance matrix used in the numerical simulations is

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \psi\mathbf{\Sigma}_{11} \\ \psi\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{22} \end{bmatrix}. \tag{5.65}$$

In this context, the intra-correlation between the state variables in $\mathbf{M}_1$ is modelled by the $\upsilon_{11}$ parameter, the intra-correlation between the state variables in $\mathbf{M}_2$ is modelled by the $\upsilon_{22}$ parameter and the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ modelled by $\psi$. Numerical analysis shows that a larger value of $\upsilon_{11}$ results in a more correlated matrix $\mathbf{M}_1$. Moreover, the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ increases with the value of $\psi$. The following lemma shows that that $\mathbf{\Sigma}$ is positive semi-definite when $\upsilon_{11} = \upsilon_{22}$.

**Lemma 5.** *Let $\psi \in [0, 1]$ and $\mathbf{A} \in S_+^M$, then the matrix $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{A} & \psi\mathbf{A} \\ \psi\mathbf{A} & \mathbf{A} \end{bmatrix} \in S_+^{2M}$.*

However, when $\upsilon_{11} = \upsilon_{22}$ the model generates matrices for which the rank of the combined matrices are equal. To be able to generate matrices for all the rank cases in Table 5.1, the model in (5.65) is used.

Note that for the model in (5.65) Lemma 5 does not apply and consequently, for some values of $\upsilon_{11}$,$\upsilon_{22}$ and $\psi$ the resulting matrix $\mathbf{\Sigma}$ is not positive semi-definite. It is worth noting that this constraint does not affect the capability of the model to generate matrices for all the rank values of interest.

Furthermore, the matrix $\mathbf{M}$ generated using the model in (5.65) is not exactly low rank. Instead, it can be well approximated by a low rank matrix. Consider the singular value decomposition of the matrix $\mathbf{M}$ given by

$$\mathbf{M} = \mathbf{USV}^T, \tag{5.66}$$

where

$$\mathbf{S} = \text{diag}\{\sigma_1, \sigma_2, ..., \sigma_{\min\{2M,N\}}\}, \tag{5.67}$$

where $\sigma_i$, for $i \in \{1, 2, ..., \min\{2M, N\}\}$ are the singular values of $\mathbf{M}$ in descending order.

Let us denote by $\tilde{\mathbf{M}}(r)$ the low rank approximation for the matrix $\mathbf{M}$ of rank $r$. Consequently, $\tilde{\mathbf{M}}(r) : \mathbb{R}^{M\times N} \to \mathbb{R}^{M\times N}$, where $\tilde{\mathbf{M}}(r)$ is given by

$$\tilde{\mathbf{M}}(r) = \mathbf{U}\mathbf{S}_r\mathbf{V}^T, \tag{5.68}$$

with

$$\mathbf{S}_r = \text{diag}\{\sigma_1, \sigma_2, ..., \sigma_r\}. \tag{5.69}$$

In the following, $r$ is defined as the minimum value for which the normalized error between the matrix $\mathbf{M}$ and the low rank approximation of rank $r$, i.e., $\tilde{\mathbf{M}}(r)$, is below $10^{-3}$, i.e.,

$$\frac{\|\mathbf{M} - \tilde{\mathbf{M}}(r)\|_2^2}{\|\mathbf{M}\|_2^2} \leq 10^{-3}. \tag{5.70}$$

Consequently, the model in (5.65) is used to generate data matrices $\mathbf{M}$ such that the low rank approximations $\tilde{\mathbf{M}}_1(r_1)$, $\tilde{\mathbf{M}}_2(r_2)$ and $\tilde{\mathbf{M}}(r)$ are of rank values depicted in Table 5.1. Moreover, the low rank approximation of the combined matrix, i.e.,$\tilde{\mathbf{M}}(r)$, is used to evaluate the numerical performance of both BSVT and SVT in exploiting the correlation between the two datasets.

Recall that $k_1$, $k_2$ and $k$ denote the number of available entries for the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ respectively. Therefore, $k$ and is given by

$$k = k_1 + k_2. \tag{5.71}$$

The locations of the available entries are sampled uniformly at random in each dataset. Note that while the recovery is done for the matrix $\tilde{\mathbf{M}}(r)$, the sampling is done independently for $\mathbf{M}_1$ and $\mathbf{M}_2$.

In the following, the numerical results for the joint recovery of two datasets are divided in three sections depending on the noise regime considered: the almost noiseless regime (SNR=50 dB), the high SNR regime (SNR=20 dB), and the medium SNR regime (SNR=10 dB). Note that the low SNR regime, i.e., SNR=0 dB, is not included as it does not provide additional insight and it increases the size of the numerical section considerably. Moreover, for each SNR regime, three ratios

between the rank of the matrices combined are considered: the equal rank scenario ($r_1 = 6$, $r_2 = 6$), the small ratio case ($r_1 = 6$, $r_2 = 9$) and the large ratio scenario ($r_1 = 6$, $r_2 = 18$). For each rank combination a high cross-correlation and a low cross-correlation case is considered. Also, the performance for each case is averaged over 20 realizations of $\Omega$.

For each noise regime, the following scenarios are addressed:

1. Equal rank: $r_1 = 6$ and $r_2 = 6$
   (a) High cross-correlation: $r = 9$
   (b) Low cross-correlation: $r = 12$
2. Small rank ratio: $r_1 = 6$ and $r_2 = 9$
   (a) High cross-correlation: $r = 10$
   (b) Low cross-correlation: $r = 14$
3. Large rank ratio: $r_1 = 6$ and $r_2 = 18$
   (a) High cross-correlation: $r = 19$
   (b) Low cross-correlation: $r = 22$

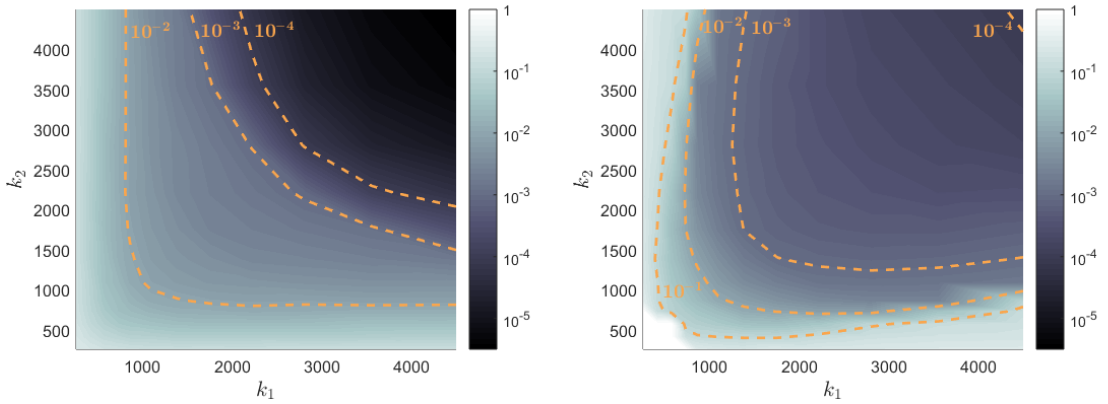
The following numerical results present a joint recovery performance comparison between the SVT algorithm and the BSVT approach. The objective is to assess how efficient are the two algorithms in exploiting cross-correlation.

### 5.6.2 Joint recovery performance in the almost noiseless regime

This section compares the numerical recovery performance for the SVT and BSVT algorithms in the almost noiseless regime in which SNR=50 dB. To that end, three ratios between the rank of the combined matrices are considered and for each ratio a highly correlated and a slightly correlated scenario is analyzed. The following section, presents the numerical results for the case in which the values of the rank for the two matrices are equal.

#### Equal rank

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 6$ and SNR=50 dB. Moreover, a high cross-correlation case is modelled using a value of $r$ below the threshold given by Theorem 2, i.e., $r = 9$ and a low cross-correlation scenario is analyzed when the value of $r$ does not satisfy the conditions of Theorem 2, i.e., $r = 12$.

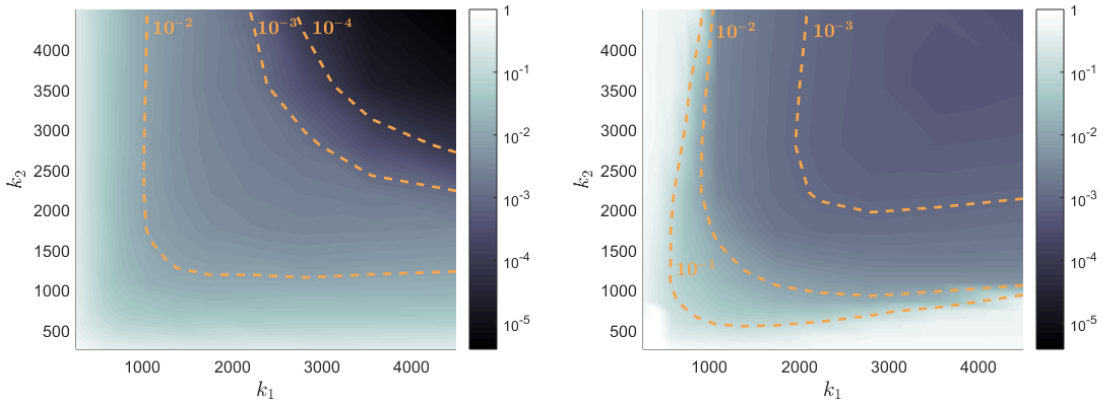



(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.16. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB.

Figure 5.16a depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB. As expected, the recovery error decreases as the number of available entries increases. Moreover, the performance does not change significantly when the ratio between the number of entries in the first matrix and the number of available entries in the second matrix increases. Interestingly, the contour lines for the $10^{-4}$ and $10^{-3}$ recovery error exhibit a similar shape to the line depicted by $\underline{k} = k_1 + k_2$ in Figure 5.1. This suggests that the BSVT algorithm successfully exploits the cross-correlation in that region and obtains a similar performance when the number of observations from one dataset varies significantly. In contrast, the contour line for $10^{-2}$ recovery error has a similar shape with the region $\mathcal{R}_4$ in Figure 5.1. This shape is equivalent to a change in recovery performance when the total number of observation is constant but the ratio between $k_1$ and $k_2$ varies. In other words, the recovery algorithm is not able to fully compensate for the lack of observations from one dataset. It is worth mentioning that in the almost noiseless scenario the BSVT algorithm is able to exploit the cross-correlation more efficiently when the ratio of observed entries is larger that 0.5.

Figure 5.16b shows the recovery performance of the SVT algorithm when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB. Note that both Figure 5.16a and Figure 5.16b have the same color map which enables a direct comparison for the recovery performance. In view of this, the BSVT algorithm outperforms the SVT algorithm when a large number of observations is available. Moreover, for the SVT algorithm the contour lines for $10^{-3}$, $10^{-2}$ and $10^{-1}$ recovery error exhibit a similar shape to the region $\mathcal{R}_4$ in Figure 5.1. This suggest that the algorithm does not exploit the cross-correlation efficiently as the recovery error changes with the ratio between $k_1$ and $k_2$ for a fixed number of observations. It is worth noting that for the SVT recovery the area with recovery error above $10^{-3}$ is larger compared with the BSVT recovery. Specifically,




(a) Recovery performance using BSVT (b) Recovery performance using SVT

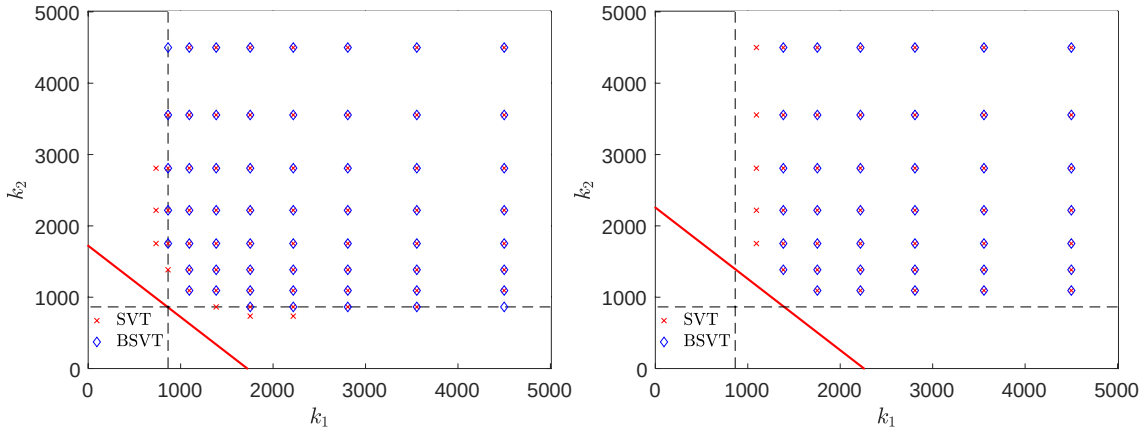
Figure 5.17. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=50 dB.

when $k_1, k_2 \in [1500, 2500]$ the SVT performs better while the BSVT achieves a smaller recovery error in the rest of the cases.

Figure 5.17a depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=50 dB. Note that the in comparison to Figure 5.16a, the contour lines are shifted towards the right side. This shows a decrease in recovery performance due to the low level of cross-correlation. As a consequence, the area with recovery error below $10^{-3}$ is smaller compared to the case in which $r = 9$. However, the shape of the contour lines is similar to the case depicted in Figure 5.16a, which suggests that BSVT algorithm is still able to obtain similar recovery performance for different ratios between $k_1$ and $k_2$ when the ratio of observed entries is above 0.6.

Figure 5.17b shows the performance of the SVT algorithm when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=50 dB. Similar to the BSVT recovery, a decrease in performance is observed for the SVT algorithm in comparison to the high cross-correlation case. Specifically, the contour lines are shifted towards larger values of $k_1$ which shows a decrease in performance caused by the lack of cross-correlation between the two datasets. The shape of the contour lines is similar to the shape of region $\mathcal{R}_4$ in Figure 5.1 which suggests that the SVT algorithm is not efficient in exploiting cross-correlation in this case. As a consequence, the recovery performance changes significantly when the ratio between $k_1$ and $k_2$ varies.

Figure 5.18a depicts the cases in which the two algorithms obtain a recovery error below $10^{-2}$ while the red line represents the theoretical limit for the joint recovery of the two datasets. Interestingly, the region that contains all the points has a similar shape to the region with probability one in Figure 5.9b. Also, the numerical results are close to the theoretical limit with the SVT algorithm performing better closer to the limit.




(a) $r_1 = 6$ and $r_2 = 6$ and $r = 9$ (b) $r_1 = 6$ and $r_2 = 6$ and $r = 12$

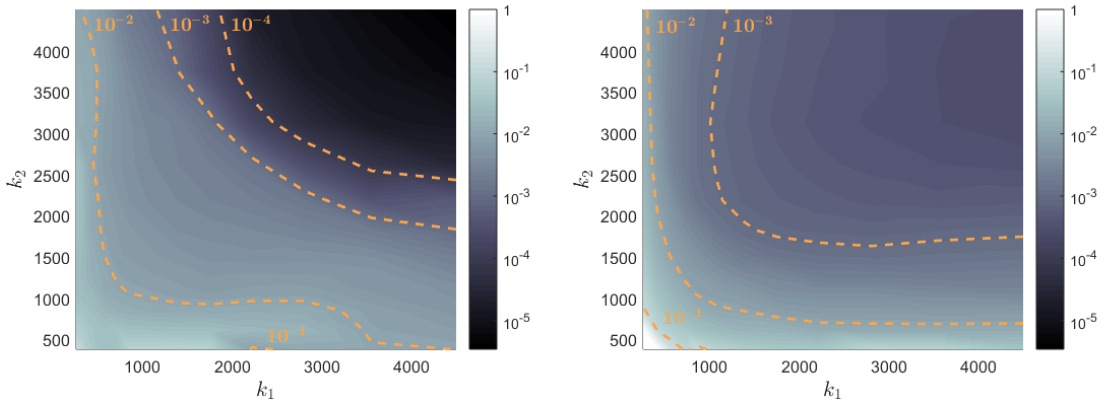
Figure 5.18. Successful joint recovery scenarios using SVT and BSVT for the equal rank case, when SNR=50 dB.

Figure 5.18b depicts the successful recovery cases in which the recovery error is below $10^{-2}$ for each algorithm. The red line represents the theoretical limit for the joint recovery of the two datasets in this particular case. Similar to the case depicted in Figure 5.18a, the region that contains all the points has a similar shape to the region with probability one in Figure 5.9b. It is worth noting that as the theoretical limit shifts towards right due to the smaller value of cross-correlation, the successful recovery cases exhibit a similar behavior. Comparing Figure 5.18b with Figure 5.18a suggests that there is a constant distance between the theoretical limit and the numerical results for the successful recovery of two datasets.

### Small rank ratio

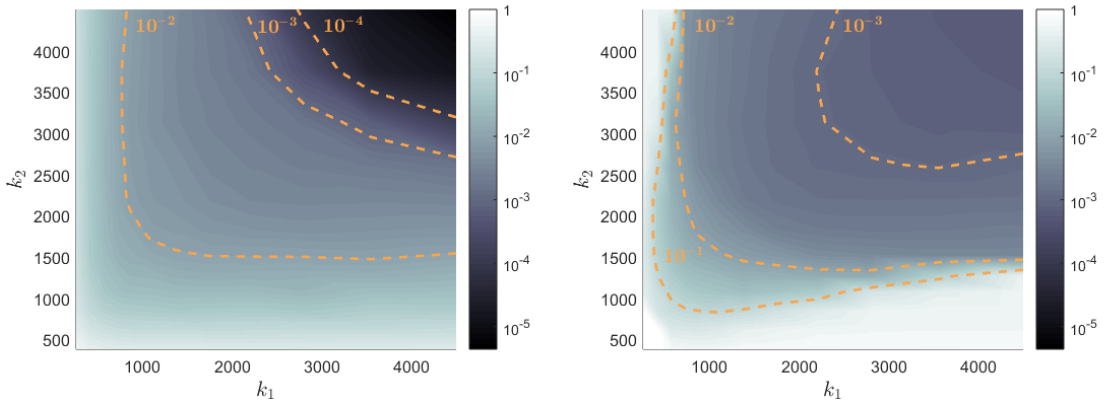
This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 9$ and SNR=50 dB. A large value of cross-correlation is modelled using a value of $r$ below the threshold given by Theorem 2, i.e. $r = 10$ and a small value of cross-correlation is analyzed when the value of $r$ does not satisfy the conditions of Theorem 2, i.e. $r = 14$.

Figure 5.19a depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=50 dB. As expected, the error decreases as the number of available entries increases. Interestingly, the contour lines for the $10^{-4}$ and $10^{-3}$ recovery error exhibit a similar shape to the line depicted by $\underline{k} = k_1 + k_2$ in Figure 5.1. This suggests that the BSVT algorithm is able to exploit the cross-correlation in that region and obtain a similar performance when the number of observations from one dataset varies significantly. In contrast, the contour line for $10^{-2}$ recovery error has a similar shape with the region $\mathcal{R}_4$ in Figure 5.1 with a slight deviation on the right side. However, with the exception of a few extreme cases the recovery error does not




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.19. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=50 dB.



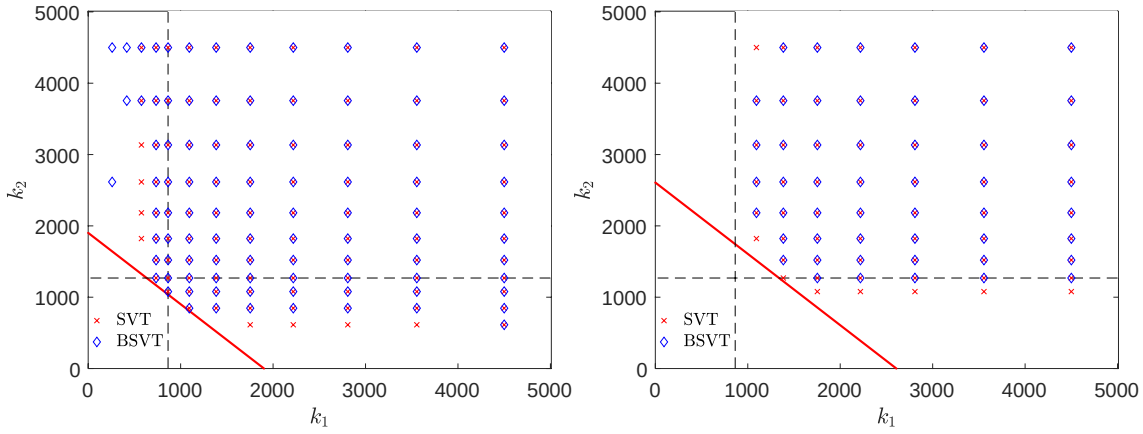

(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.20. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=50 dB.

change significantly when the total number of observations is constant but the ratio between $k_1$ and $k_2$ varies. In view of this, the BSVT algorithm is able to exploit the cross-correlation in the small rank ratio case.

In Figure 5.19b the performance of the SVT algorithm is depicted for the case in which $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=50 dB. Similar to the equal rank case, the BSVT algorithm outperforms the SVT-based recovery for most of the values of $k_1$ and $k_2$ considered. Also, the shape of the contour lines for $10^{-3}$ and $10^{-2}$ recovery error is similar to the shape of region $\mathcal{R}_4$ in Figure 5.1 which suggests that the SVT algorithm is not efficient in exploiting cross-correlation.

Figure 5.20a shows the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=50 dB. In comparison with the case in which $r = 10$, the area with recovery error below $10^{-3}$ is smaller. This suggests a decrease in recovery




(a) $r_1 = 6$ and $r_2 = 9$ and $r = 10$ (b) $r_1 = 6$ and $r_2 = 9$ and $r = 14$

Figure 5.21. Successful joint recovery scenarios using SVT and BSVT for the small rank ratio case, when SNR=50 dB.

performance due to the small value of cross-correlation modelled in this scenario. In addition, the shape of the contour line for $10^{-2}$ recovery error suggests that for some ratios between $k_1$ and $k_2$, a decrease in recovery performance is observed compared to the highly correlated case in which $r = 10$.

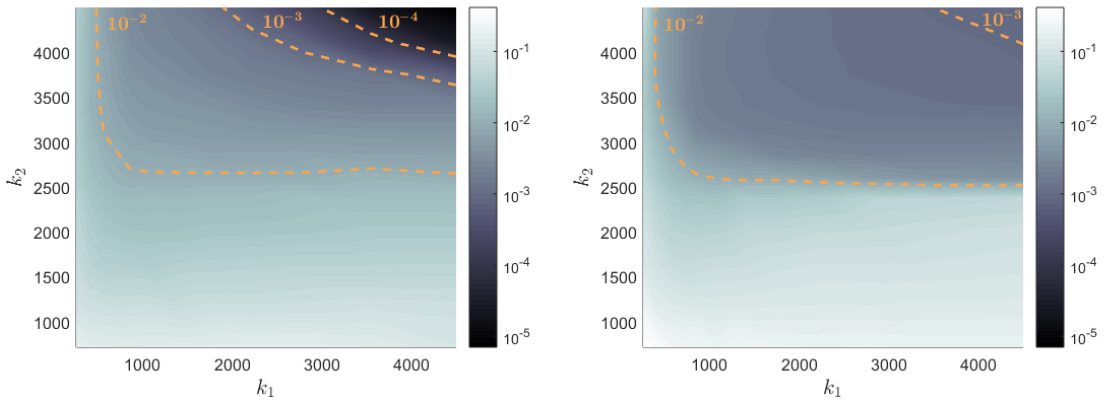
Figure 5.20b depicts the performance of the SVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=50 dB. In comparison with the BSVT algorithm there is a significant decrease in recovery performance for the cases in which the number of observations from one dataset is below 1000. Also, the shape of the contour lines indicates that the SVT approach is not able to exploit the cross-correlation in this case.

Figure 5.21a shows the successful recovery cases in which the error is below $10^{-2}$ for each of the two algorithms. Also, the red line depicts the theoretical limit for the joint recovery of the two datasets in this particular case. Interestingly, the numerical results in this case are close to the theoretical limit.

Figure 5.21b shows the successful recovery cases in which the error is below $10^{-2}$ for each of the two algorithms. The red line depicts the theoretical limit for the joint recovery of the two datasets in this particular case. Similar to the highly correlated case depicted in Figure 5.21a, the numerical results in this case are close to the theoretical limit. Moreover, the shift towards the right side of the theoretical limit is observed in the numerical results as well.

### Large rank ratio

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 18$ and SNR=50 dB. The high cross-correlation case is modelled using $r = 19$ and the low cross-correlation scenario is analyzed when




(a) Recovery performance using BSVT (b) Recovery performance using SVT

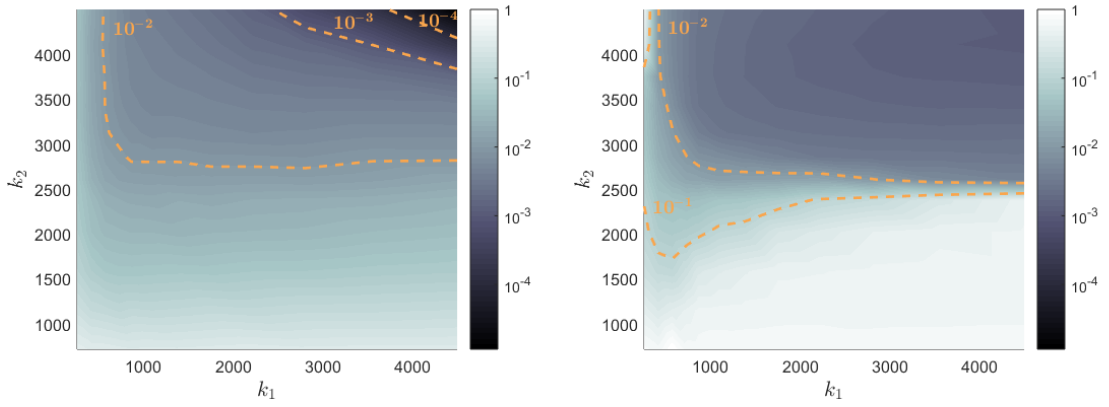
Figure 5.22. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=50 dB.

$r = 22$. Note that the conditions of Theorem 2 are not satisfied even for the highly correlated datasets scenario.

Figure 5.22a shows the recovery performance for the BSVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=50 dB. In comparison with the other highly correlated scenarios depicted in Figure 5.16a and Figure 5.19a the area with recovery error below $10^{-3}$ is smaller. Not surprisingly, the ratio between the rank of the matrices combined has a significant impact on the rank of the combined matrix and ultimately on the joint recovery performance. However, the shape of the contour lines for $10^{-4}$ and $10^{-3}$ recovery error suggests there is no significant change in recovery performance for different ratios between $k_1$ and $k_2$ when a larger proportion of the entries are available. In contrast, when the number of observations from the second dataset is below 2500, the recovery error is always above $10^{-2}$. This is because the rank of the second matrix is significantly larger compared to the rank of the first matrix.

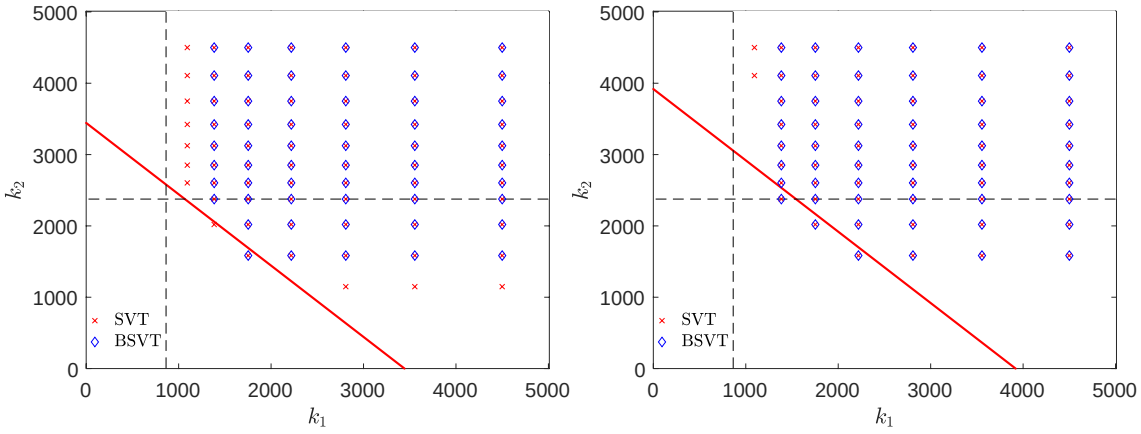
Figure 5.22b depicts the recovery performance of the SVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=50 dB. In comparison with the BSVT recovery depicted in Figure 5.22a, the area with recovery error below $10^{-3}$ is significantly smaller while the area with recovery error below $10^{-2}$ is the same. Also, when the number of observations from the second dataset is below 2500, the recovery error is always above $10^{-2}$ which shows the impact of the rank of the second matrix.

Figure 5.23a illustrates the recovery error for the BSVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=50 dB. As expected, a decrease in the area with recovery error below $10^{-3}$ is observed compared to the highly correlated case depicted in Figure 5.22a. This observation is consistent with the other numerical results presented in this section.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.23. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=50 dB.

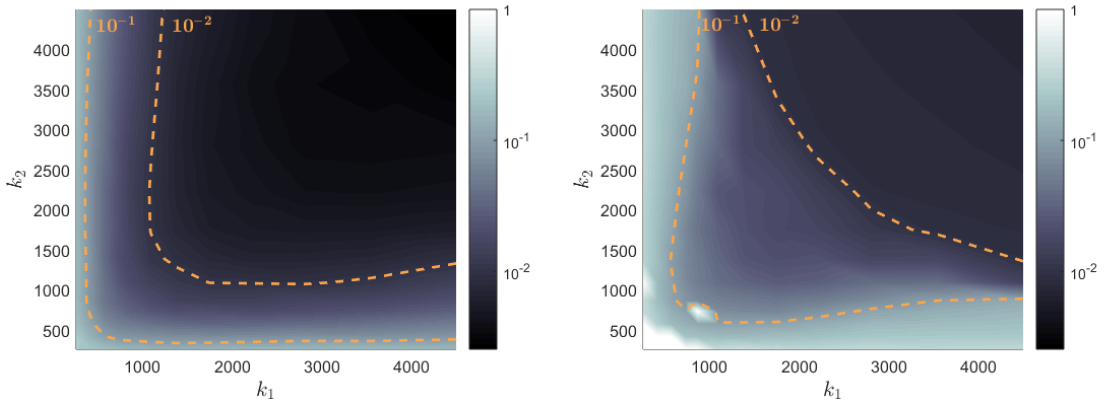



(a) $r_1 = 6$ and $r_2 = 18$ and $r = 19$ (b) $r_1 = 6$ and $r_2 = 18$ and $r = 22$

Figure 5.24. Successful joint recovery scenarios using SVT and BSVT for the large rank ratio case, when SNR=50 dB.

Figure 5.23b depicts the performance of the SVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=50 dB. A decrease in performance is also observed in this case compared to the highly correlated case depicted in Figure 5.22b. Specifically, when the number of available entries from the second dataset is below 2500 the recovery error increases above $10^{-1}$. In contrast to the BSVT approach, the SVT algorithm is not able to achieve a recovery error below $10^{-3}$ in this case.

Figure 5.24a illustrates a comparison between the theoretical limit and the numerical results using SVT and BSVT. Interestingly, the two recovery algorithm are able to perform close to the theoretical limit in the almost noiseless scenario. This observation is consistent with the previous results presented in this section. Also, the shape of the region that contains all the cases with successful recovery is similar to the region with probability one depicted in Figure 5.12b.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.25. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=20 dB.

Figure 5.24b shows the cases in which the recovery performance is below $10^{-2}$ for each algorithm. Admirably, both approaches perform close to the theoretical limit represented by the red line. It is also worth noting that as the rank of the combined matrix increases and the theoretical limit shifts towards right, the numerical results exhibit the same behavior.

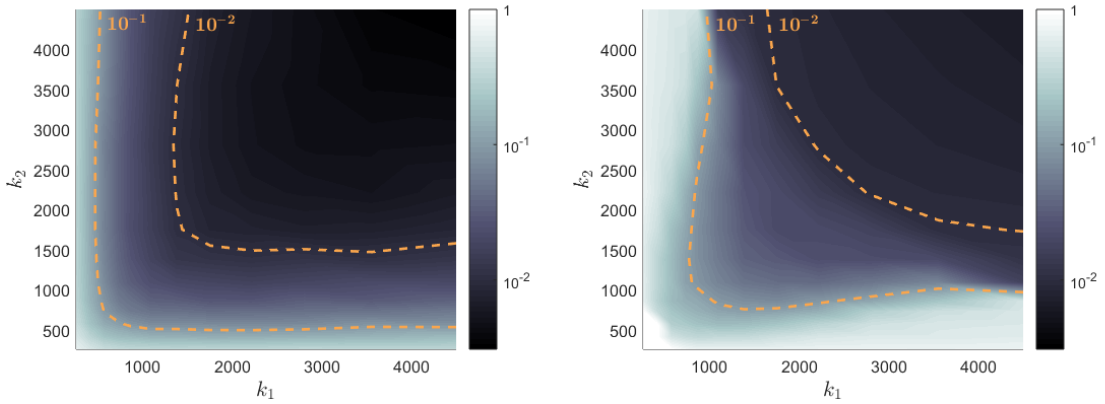
### 5.6.3 Joint Recovery performance in the high SNR regime

This section shows the numerical recovery performance for the SVT and BSVT algorithms in the high SNR regime in which SNR=20 dB. Similar to the almost noiseless case, three ratios between the rank of the combined matrices are considered and for each ratio a highly correlated and a slightly correlated scenario is illustrated. The following section, presents the numerical results for the case in which the values of the rank for the two matrices are equal.

#### Equal rank

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 6$ and SNR=20 dB. Moreover, a large value of cross-correlation is modelled by $r = 9$ and a small value of cross-correlation scenario is analyzed when $r = 12$.

Figure 5.25a depicts the BSVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=20 dB. A significant decrease in performance is observed compared to the almost noiseless case illustrated in Figure 5.16a. However, this is expected and consisted with the numerical results presented in Section 4.4. Similar to the SNR=50 dB scenario, the recovery performance does not change significantly with the ratio between $k_1$ and $k_2$ when more than half of the observations are available. On the




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.26. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=20 dB.

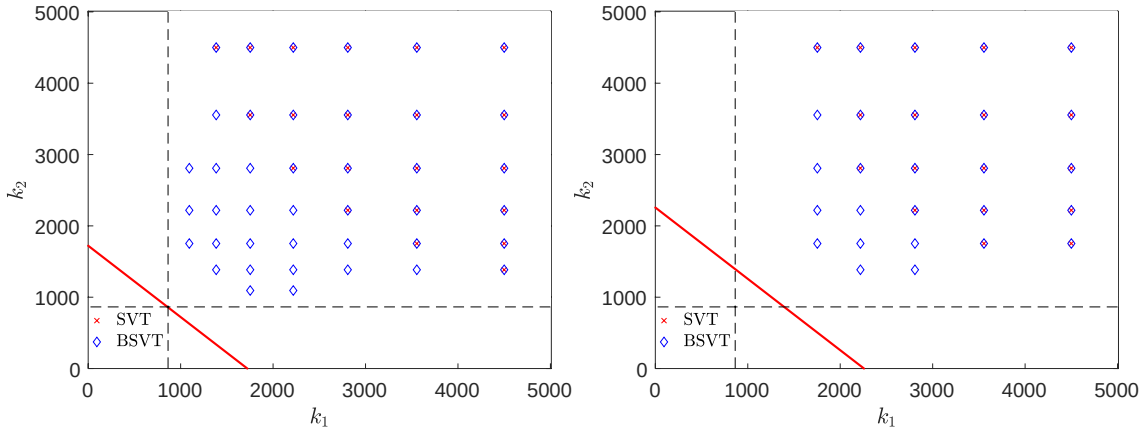
other hand, the shape of the contour lines for recovery error $10^{-2}$ and $10^{-1}$ suggests that when less than half of the observations are available, the performance changes with the ratio between $k_1$ and $k_2$.

Figure 5.25b shows the SVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=20 dB. Comparing Figure 5.25a and Figure 5.25b shows that the BSVT algorithm outperforms the SVT recovery in this particular case. However, the contour line for recovery error of $10^{-2}$ suggests that when the ratio of observed entries is above 0.6, the performance does not change significantly with the ratio between $k_1$ and $k_2$.

Figure 5.26a depicts the BSVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=20 dB. In comparison to the case in which $r = 9$ the area with recovery error below $10^{-2}$ is smaller. Also, the recovery performance does not change significantly with the ratio between $k_1$ and $k_2$ when more than half of the observations are available.

Figure 5.26b shows the SVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=20 dB. A decrease in performance is observed in this case compared to the case in which $r = 9$ depicted in Figure 5.25b. In particular, the loss in recovery performance is observed for the cases in which less than half of the entries are observed. As expected, the BSVT algorithm outperforms the SVT recovery for this particular case as well. In terms of exploiting the cross-correlation, the recovery error does not change significantly when the ratio of observed entries is above 0.6.

Figure 5.27a illustrates the cases in which the recovery performance is below $10^{-2}$ for each algorithm when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=20 dB. In contrast to the almost noiseless regime, only the BSVT approach performs close to the theoretical limit depicted by the red line. It is worth noting that for each algorithm the shape of the region that contains all the cases with successful recovery is similar to the




(a) $r_1 = 6$ and $r_2 = 6$ and $r = 9$ (b) $r_1 = 6$ and $r_2 = 6$ and $r = 12$

Figure 5.27. Successful joint recovery scenarios using SVT and BSVT for the equal rank case, when SNR=20 dB.

region of probability one in Figure 5.9b. However, the area corresponding to the BSVT algorithm is considerable larger than that of the SVT approach and it is also closer the the theoretical limit.

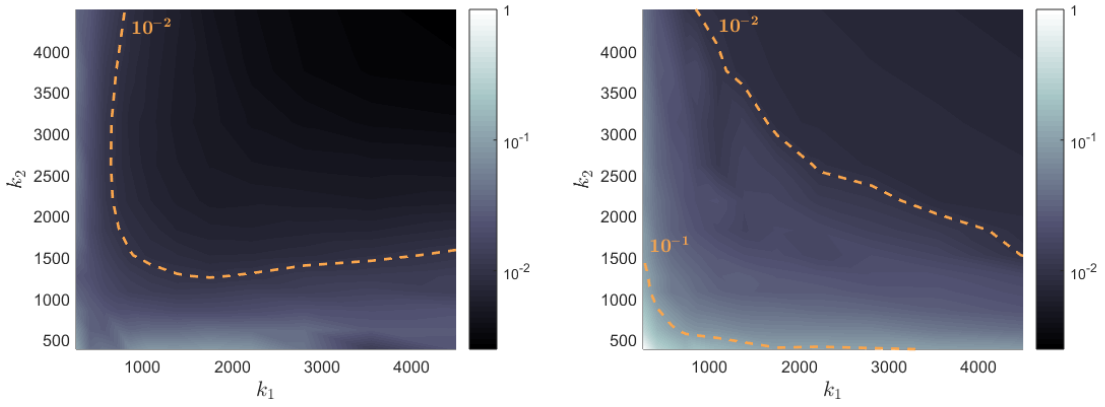
Figure 5.27b shows the successful recovery cases for SVT and BSVT when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=20 dB. The smaller value of cross-correlation in this case shifts the theoretical limit in this case towards the right side. Interestingly, the save behavior is observed for points corresponding to the cases in which the recovery is successful. Also, the BSVT algorithm performs close to the theoretical limit in this. This observation is consistent with the numerical results depicted in Figure 5.27a.

### Small rank ratio

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 9$ and SNR=20 dB. Moreover, a large value of cross-correlation is modelled by $r = 10$ and a small value of cross-correlation scenario is analyzed when $r = 14$.

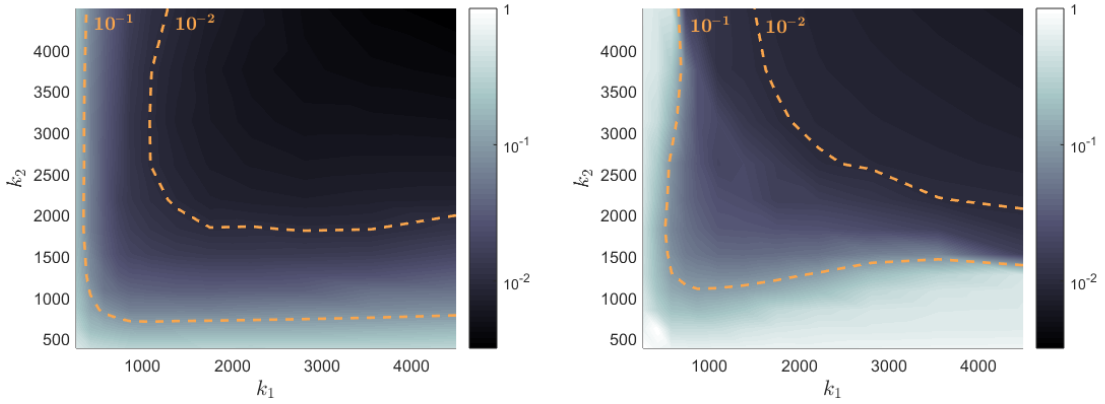
Figure 5.28a shows the recovery error measured for the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=20 dB. Interestingly, the shape of the contour line with recovery of $10^{-2}$ suggests that when less than half of the entries are observed the performance changes with the ratio between $k_1$ and $k_2$. This observation is consistent with the numerical results presented above. In contrast, when more than half of the observations are available, the recovery performance does not change significantly.

Figure 5.28b depicts the performance of the SVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=20 dB. A decrease in recovery performance is observed when compared with the BSVT recovery. However, the shape of the contour line with




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.28. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=20 dB.



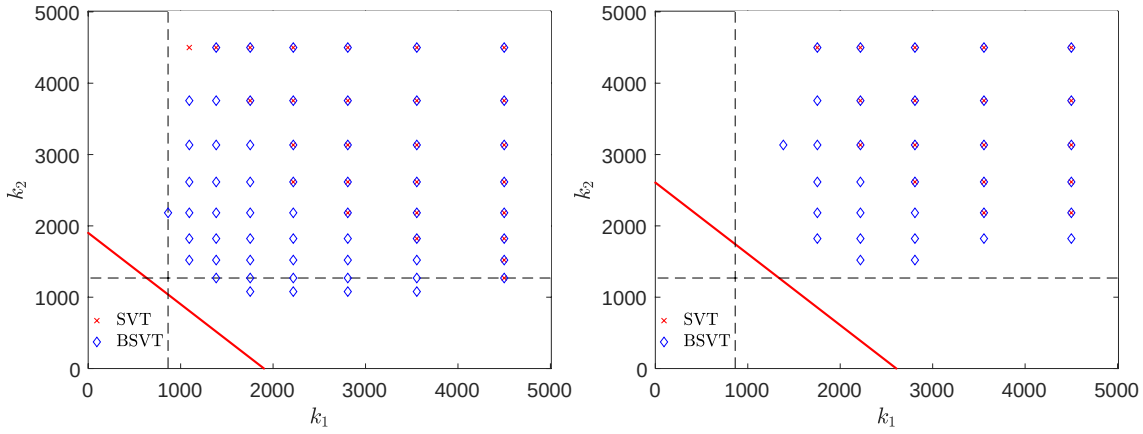

(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.29. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=20 dB.

recovery error of $10^{-2}$ suggests that the SVT algorithm is able to exploit the cross-correlation when the ratio of the observed entries is above 0.55.

Figure 5.29a shows the performance of the BSVT approach when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=20 dB. A decrease in recovery performance is observed in comparison to the case in which the rank of the combined matrix is 10. Precisely, in this case there exists a region with recovery error above $10^{-1}$ and the area with recovery error below $10^{-2}$ is smaller. However, the BSVT algorithm is still able to exploit the cross-correlation when more than half of the entries are observed.

In Figure 5.29b the performance of the SVT algorithm is depicted for the case in which $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=20 dB. As expected, the SVT approach is outperformed by the BSVT algorithm in the high SNR regime. A decrease in the number of cases for which the performance does not change with the ratio between




(a) $r_1 = 6$ and $r_2 = 9$ and $r = 10$ (b) $r_1 = 6$ and $r_2 = 9$ and $r = 14$

Figure 5.30. Successful joint recovery scenarios using SVT and BSVT for the small rank ratio case, when SNR=20 dB.

$k_1$ and $k_2$ is also observed compared with the BSVT recovery. In addition, when the number of available observation from one of the datasets is below 1000 the performance of the SVT recovery decreases significantly in this case.

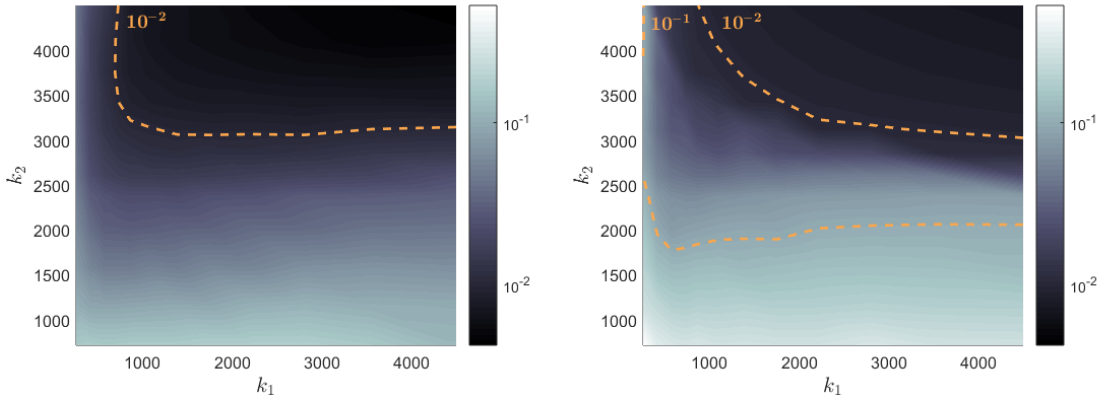
Figure 5.30a illustrates the successful recovery cases for SVT and BSVT when $r_1 = 6$, $r_2 = 6$, $r = 10$ and SNR=20 dB. Similar to the numerical results for equal rank values, the BSVT algorithm performs closer to the theoretical limit compared with the SVT approach.

Figure 5.30b depicts the successful recovery cases for SVT and BSVT when $r_1 = 6$, $r_2 = 6$, $r = 14$ and SNR=20 dB. In comparison with the highly correlated case depicted in Figure 5.30a, a shift towards right for both the theoretical limit and the numerical results is observed. Also, the BSVT recovery performs closed to the theoretical limit compared with the SVT approach.

### Large rank ratio

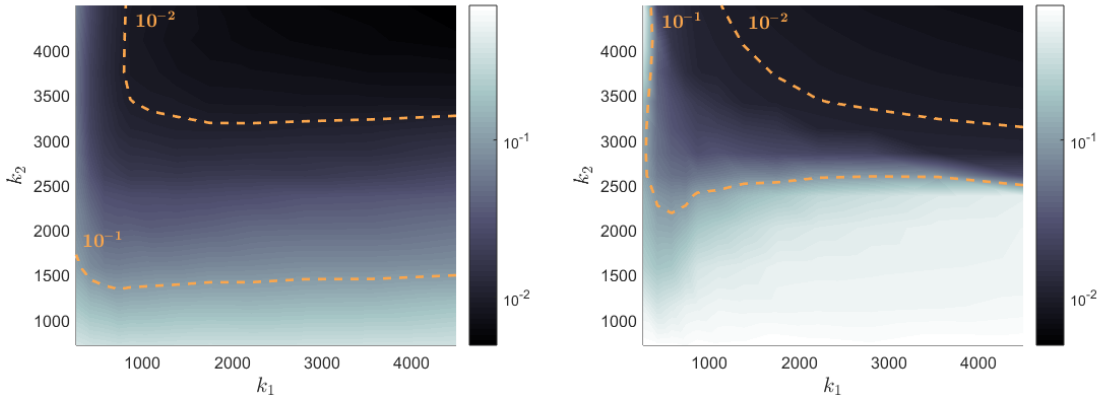
This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 18$ and SNR=20 dB. Moreover, a large value of cross-correlation is modelled by $r = 19$ and a small value of cross-correlation scenario is analyzed when $r = 22$. Note that the conditions of Theorem 2 are not satisfied even for the highly correlated datasets scenario.

Figure 5.31a shows the recovery performance for the BSVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=20 dB. Interestingly, the large value for the rank of the second matrix significantly impacts the performance of the algorithm when the number of observations from the second dataset is below 2500. This is also observed in the corresponding numerical results for the almost noiseless regime. The shape of the contour line for recovery error of $10^{-2}$ suggests that the cross-correlation is not




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.31. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=20 dB.



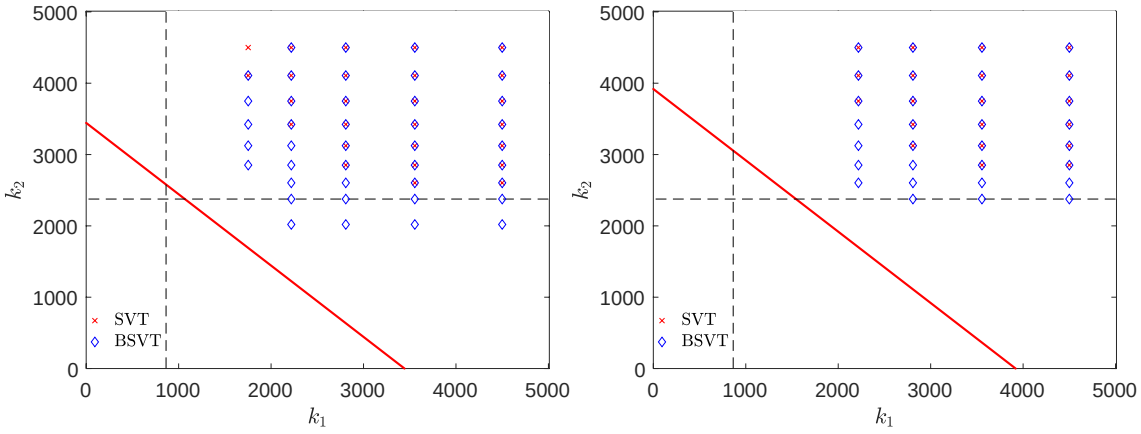

(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.32. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=20 dB.

exploited in most of the cases. In fact, the recovery error does not change with the ratio between $k_1$ and $k_2$ only when more than 70% of the observations are available.

Figure 5.31b depicts the performance of the SVT approach when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=20 dB. A decrease in recovery performance is observed compared to the BSVT algorithm. However, the number of cases in which the performance does not change with the ratio between $k_1$ and $k_2$ is the same for both recovery methods.

Figure 5.32a shows the recovery performance for the BSVT algorithm when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=20 dB. A decrease in recovery performance is observed in this case compared with the one depicted in Figure 5.31a. Precisely, for a smaller value of cross-correlation there exists a region with recovery error above $10^{-1}$.




(a) $r_1 = 6$ and $r_2 = 18$ and $r = 19$ (b) $r_1 = 6$ and $r_2 = 18$ and $r = 22$

Figure 5.33. Successful joint recovery scenarios using SVT and BSVT for the large rank ratio case, when SNR=20 dB.

Figure 5.32b illustrates the performance of the SVT algorithm in the high SNR regime when $r_1 = 6$, $r_2 = 18$ and $r = 22$. In comparison to the highly correlated case depicted in Figure 5.31b, the recovery error increased significantly when the number of observations from the second dataset is below 2500.

Figure 5.33a illustrates the cases in which the recovery error is below $10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=20 dB. Consistent with the previous results in this section, the BSVT algorithm performs closer to the theoretical limit than the SVT approach.

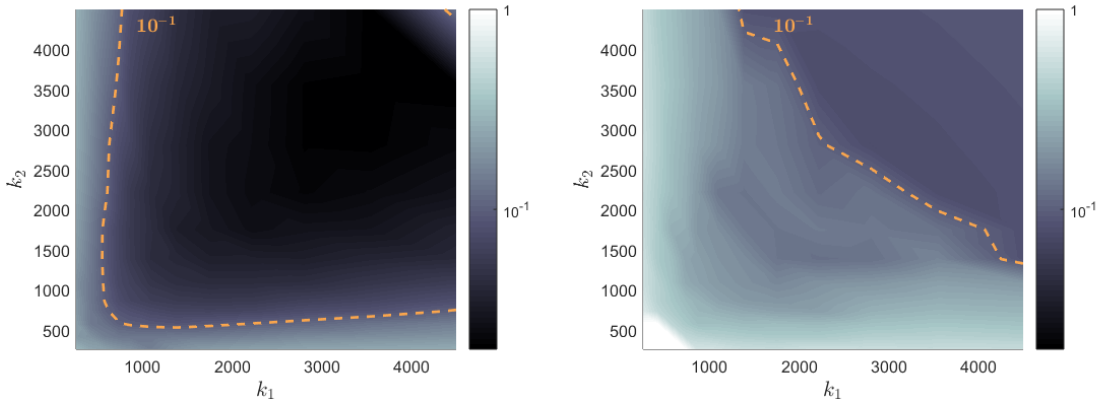
Figure 5.33b depicts the cases in which the recovery error is below $10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=20 dB. A shift towards right is observed compared to the highly correlated case in which $r = 19$. The BSVT algorithm is performing closed to the theoretical limit when compared with the SVT recovery.

### 5.6.4 Joint Recovery performance in the medium SNR regime

This section shows the numerical recovery performance for the SVT and BSVT algorithms in the medium SNR regime in which SNR=10 dB. Similar to the previous noise regimes, three ratios between the rank of the combined matrices are considered and for each ratio a highly correlated and a slightly correlated scenario is illustrated. The following section, presents the numerical results for the case in which the values of the rank for the two matrices are equal.

#### Equal rank

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 6$ and SNR=10 dB. Moreover, a large value of




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.34. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=10 dB.

cross-correlation is modelled by $r = 9$ and a small value of cross-correlation scenario is analyzed when $r = 12$.

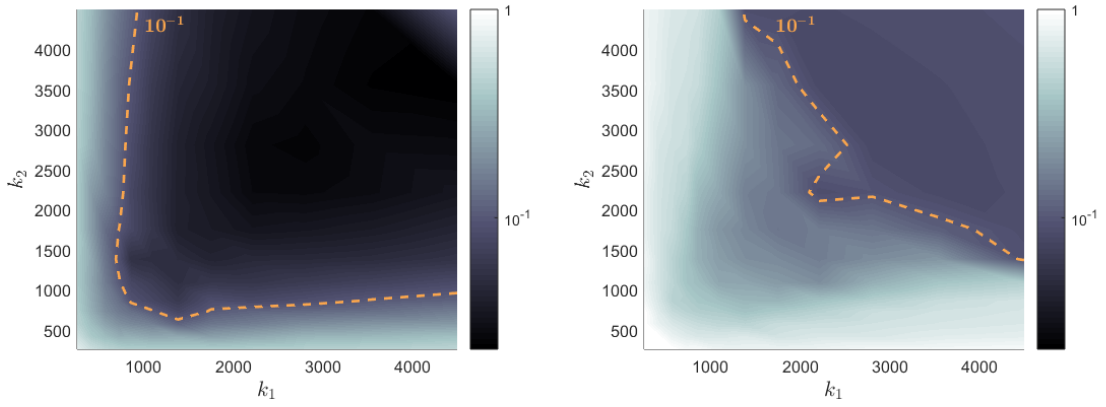
Figure 5.34a depicts the BSVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=10 dB. The shape of the contour line for the recovery error of $10^{-1}$ suggests that when less than half of the entries are observed, the performance varies with the ratio between $k_1$ and $k_2$. In contrast, when the ratio of observed entries is above 0.5, the recovery performance does not change significantly with the ratio between the number of available entries from the two datasets.

Figure 5.34b shows the SVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=10 dB. Clearly, the SVT algorithm is outperformed by the BSVT recovery in this case. This observation is consistent with the analysis in Chapter 4. Also the ratio of observed entries above which the performance does not change with the ratio between $k_1$ and $k_2$ is 0.6 for the SVT recovery. This suggests that the SVT is not as efficient as the BSVT approach in exploiting the cross-correlation in this noise regime.

Figure 5.35a depicts the BSVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=10 dB. In comparison with the highly correlated depicted in Figure 5.34a, a decrease in recovery performance is noticeable due to the shift towards right of the contour line.

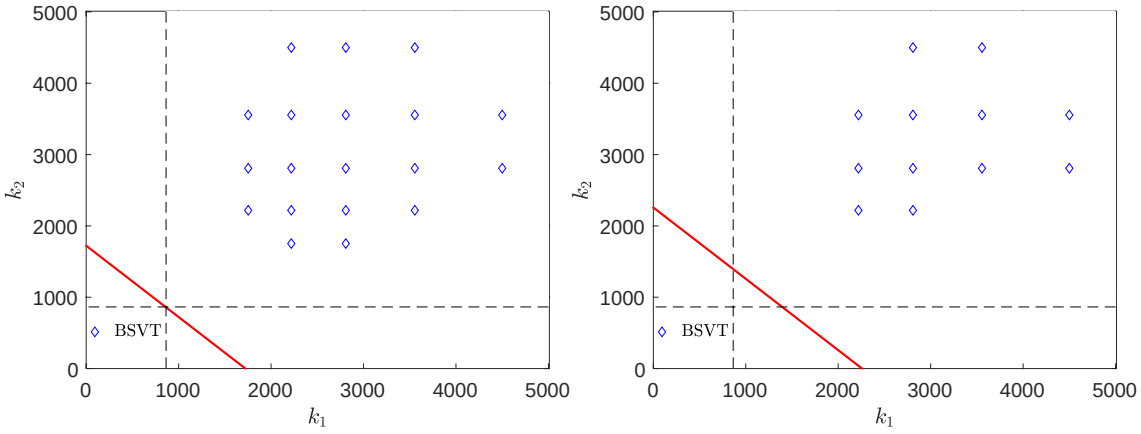
Figure 5.35b shows the SVT recovery performance when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=10 dB. Similar to the high SNR regime, the performance of the recovery is decreases significantly when less than half of the entries are available. Also, the ratio of observed entries above which the performance does not change with the ratio between $k_1$ and $k_2$ is 0.6 for the SVT recovery while for the BSVT algorithm is 0.5.

Figure 5.36a illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=10 dB.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.35. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=10 dB.

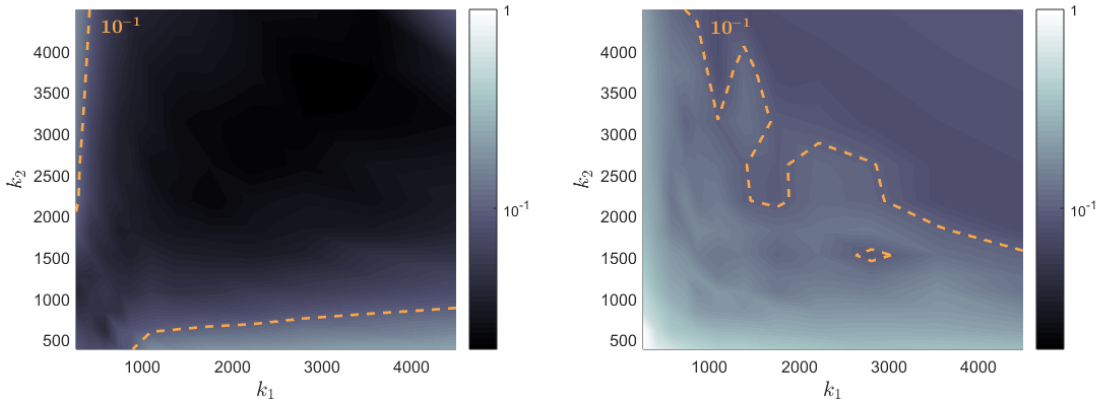



(a) $r_1 = 6$ and $r_2 = 6$ and $r = 9$ (b) $r_1 = 6$ and $r_2 = 6$ and $r = 12$

Figure 5.36. Successful joint recovery scenarios using SVT and BSVT for the equal rank case, when SNR=10 dB.

Interestingly, only the BSVT algorithm is able to successfully recovery the missing data in the medium SNR regime. However the performance of the BSVT approach is not as close to the theoretical limit as in the almost noiseless regime.

Figure 5.36b illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 6$, $r = 12$ and SNR=10 dB. Similar to the highly correlated case, only the BSVT algorithm is able to successfully recovery the missing data in the medium SNR regime. A decrease in the number of cases with successful recovery is observed compared with the results depicted in Figure 5.36a.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.37. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=10 dB.

### Small rank ratio

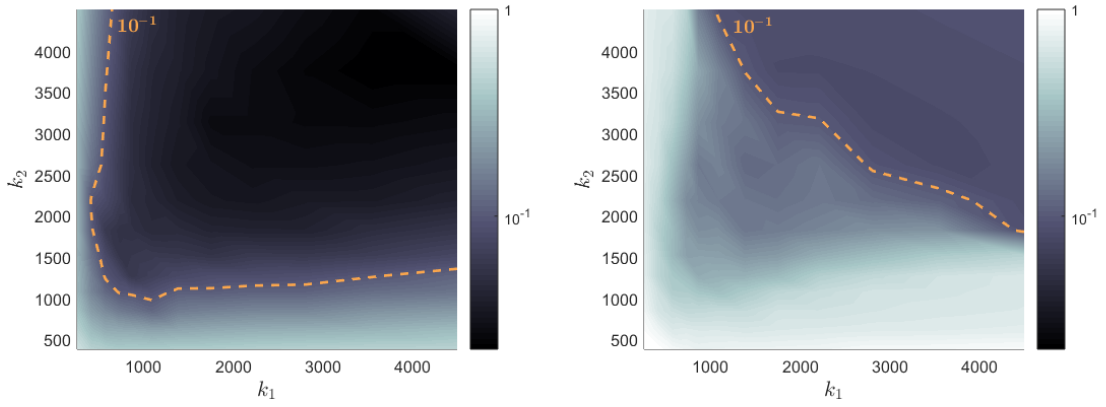
This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 9$ and SNR=10 dB. Moreover, a large value of cross-correlation is modelled by $r = 10$ and a small value of cross-correlation scenario is analyzed when $r = 14$.

Figure 5.37a depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=10 dB. Interestingly, the recovery performance is below $10^{-1}$ for almost all the cases as shown by the contour line. Also, the performance does not change significantly as the ratio between $k_1$ and $k_2$ varies for the same total number of observed entries.

Figure 5.37b shows the recovery performance for the SVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=10 dB. Clearly, the BSVT algorithm outperforms the SVT recovery in this scenario. Also, the contour line shows that the SVT algorithm is not robust and the recovery performance varies significantly for a small change in the number of observed entries.

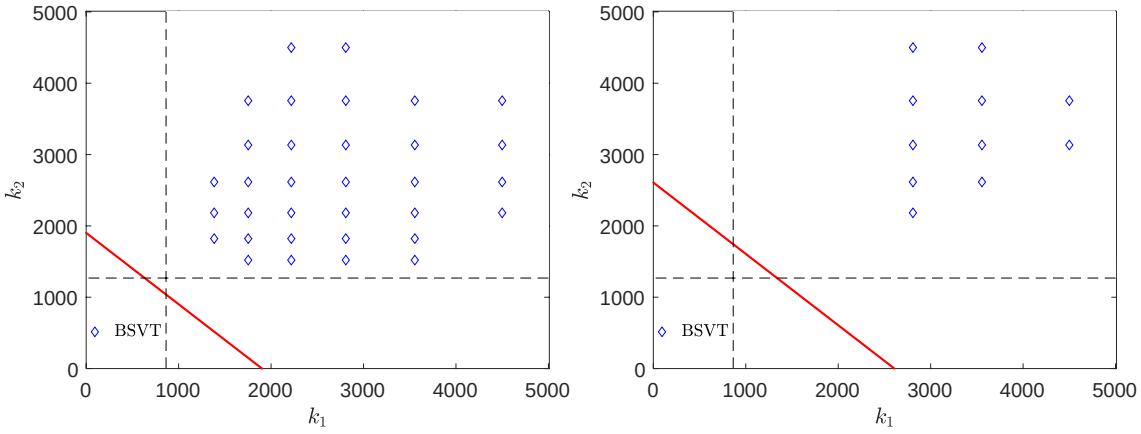
Figure 5.38a describes the recovery error for the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=10 dB. Interestingly, the performance in this case is similar to the high SNR regime depicted in Figure 5.29a. The lower level of cross-correlation modeled through a larger value of $r$ affects the recovery performance especially in the cases in which the ratio between $k_1$ and $k_2$ is large.

Figure 5.38b shows the recovery performance for the SVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=10 dB. A significant decrease in performance is observed in comparison with the BSVT recovery especially for the cases in which the ratio between $k_1$ and $k_2$ is large. This suggests that the SVT algorithm is not able to exploit the cross-correlation in a medium SNR regime.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.38. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=10 dB.

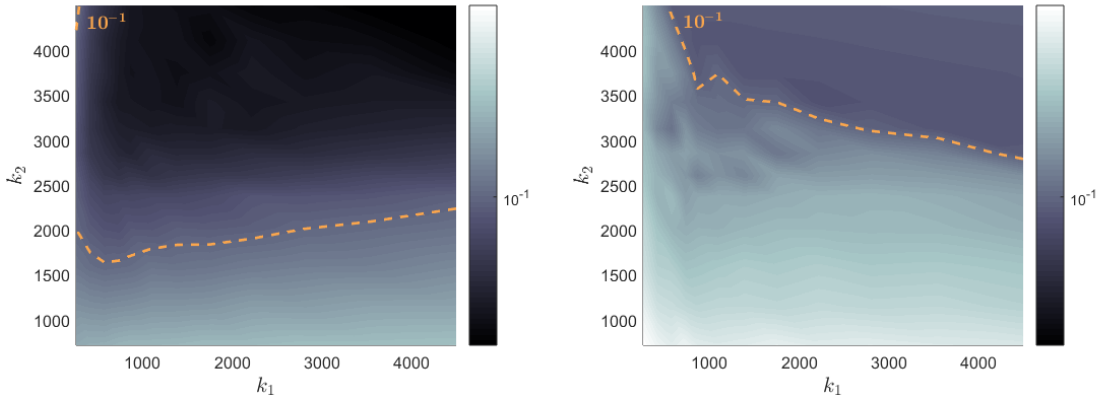



(a) $r_1 = 6$ and $r_2 = 9$ and $r = 10$ (b) $r_1 = 6$ and $r_2 = 9$ and $r = 14$

Figure 5.39. Successful joint recovery scenarios using SVT and BSVT for the small rank ratio case, when SNR=10 dB.

Figure 5.39a illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=10 dB. Similar to the previous results in the medium SNR regime, only the BSVT algorithm is able to successfully recover the missing entries.

Figure 5.39b illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 9$, $r = 14$ and SNR=10 dB. A decrease in the number of cases with successful recovery is noticeable compared with the highly correlated case depicted in Figure 5.39a. Still, only the BSVT algorithm is able to provide a recovery error below the threshold.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.40. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=10 dB.

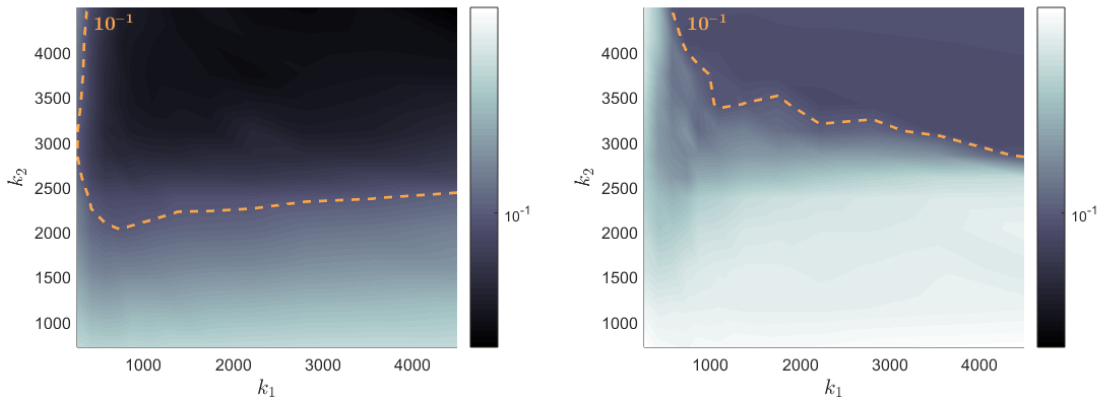
### Large rank ratio

This section compares the numerical recovery performance for the SVT and BSVT algorithms, when $r_1 = 6$ and $r_2 = 18$ and SNR=10 dB. Moreover, a large value of cross-correlation is modelled by $r = 19$ and a small value of cross-correlation scenario is analyzed when $r = 22$. Note that the conditions of Theorem 2 are not satisfied even for the highly correlated datasets scenario.

Figure 5.40a depicts the performance of the BSVT algorithm in the medium SNR regime when $r_1 = 6$, $r_2 = 18$ and $r = 19$. Interestingly, the distribution of the recovery error in this case is similar to the one in the high SNR regime. Note that the color map varies across different noise regimes. However, the decrease in recovery performance when less than 2500 observations are available from the second dataset is noticeable across all noise regimes.

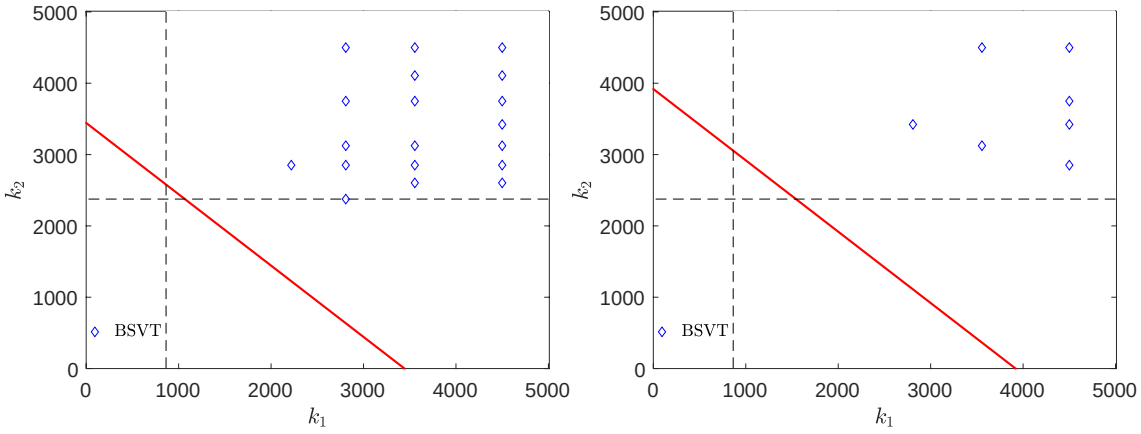
Figure 5.40b shows the recovery performance for the SVT approach when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=10 dB. As expected the BSVT algorithm outperforms the SVT recovery in the medium SNR regime. In addition, the same decrease in performance for values of $k_2$ below 2500 is observed in this case.

Figure 5.41a shows the recovery performance for the BSVT approach when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=10 dB. The distribution of the recovery error in this case is similar to the corresponding results for the high SNR regime. Also, a decrease in the recovery performance is observed compared with the highly correlated case in Figure 5.40a. For the SVT recovery depicted in Figure 5.41b a more noteworthy decrease in performance is observed when the number of observed entries from the second dataset in below 2500. This suggests that the SVT algorithm is less efficient in exploiting cross-correlation in the medium SNR regime.




(a) Recovery performance using BSVT (b) Recovery performance using SVT

Figure 5.41. Joint recovery error measured by NMSE, when $r_1 = 6$, $r_2 = 18$, $r = 22$ and SNR=10 dB.




(a) $r_1 = 6$ and $r_2 = 18$ and $r = 19$ (b) $r_1 = 6$ and $r_2 = 18$ and $r = 22$

Figure 5.42. Successful joint recovery scenarios using SVT and BSVT for the large rank ratio case, when SNR=10 dB.

Figure 5.42a illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 18$, $r = 19$ and SNR=10 dB. Similar to the previous results in the medium SNR regime, only the BSVT algorithm is able to successfully recover the missing entries and the performance is not as close to the theoretical limit as in the almost noiseless regime.

Figure 5.42b illustrates the cases in which the recovery error is below $3 \times 10^{-2}$ for both the SVT and the BSVT recovery when $r_1 = 6$, $r_2 = 19$, $r = 22$ and SNR=10 dB. A decrease in the number of cases with successful recovery is noticeable compared with the highly correlated case depicted in Figure 5.42a. Still, only the BSVT algorithm is able to provide a recovery error below the threshold.

## 5.7 Conclusions

The fundamental limits for joint recovery of multiple matrices with missing entries have been characterized in terms of the individual and combined rank. Precisely, when two datasets are jointly recovered the rank of the combined matrix is bounded based on the individual rank of the two components. Theoretical conditions under which the number of observations required by the joint recovery is smaller than what is necessary for the independent recovery are derived. A graphical representation of the recovery regions for the independent and joint settings is included to highlight the larger number of cases in which the joint recovery is feasible.

For datasets that are sampled uniformly, the probability of joint recovery is characterized and compared to the probability of recovery in the independent case. Using the new probabilistic description of the joint recovery, the problem of minimizing the sampling cost is formulated as an optimization problem for both the independent and the joint recovery cases. The numerical comparison shows that in most of the cases the joint recovery induces a smaller cost for the acquisition process than that achieved by the independent recovery of two datasets.

To facilitate the numerical analysis of the joint recovery performance, a mathematical model for generating correlated datasets is proposed. In this setting, different rank values are assessed using synthetic data and the performance of both SVT and BSVT algorithms is numerically evaluated.

Numerical results show that the cross-correlation between different datasets can be successfully exploited in a missing data recovery context. Moreover, when the number of available entries from one dataset is small, the joint recovery successfully estimates the missing entries by leveraging the information provided by the dataset with fewer missing entries. The BSVT algorithm outperforms the SVT approach in the joint recovery context. Remarkably, the performance of BSVT recovery matches the geometry of the regions depicted by the fundamental limits which suggests that BSVT is indeed better suited to exploit the cross-correlation between datasets. In addition, BSVT is more robust across different sampling regimes for the datasets combined when compared to the SVT algorithm. In view of this, the robustness of the new algorithm proposed in Chapter 4 extends to the joint recovery setting.

# Chapter 6

# Conclusions and future work

This thesis addressed the problem of recovering missing data in high-dimensional datasets generated by different subsystems of an urban system. These components include the traffic monitoring system, the low voltage electricity grid or the air quality monitoring system, among others, and their efficient management requires timely and accurate data.

For the low voltage distribution system, information theoretic results are used to characterize the fundamental limits of the recovery process by defining the optimal performance theoretically attainable by any missing data estimator. This enables the benchmarking of different state-of-the-art recovery methods and identifies the scenarios in which their performance is largely suboptimal. In this context, the LMMSE estimator and the SVT algorithm are evaluated and their shortcomings are identified in a practical setting. Since the applicability of these methods is limited in realistic scenarios, a new algorithm is proposed to address their shortcomings.

The new algorithm combines Bayesian estimation with the SVT algorithm to enable the optimization of the soft-thresholding parameter at each iteration. Moreover, the low computational cost of the SVT algorithm and the optimality of the LMMSE estimator for datasets that follow a multivariate Gaussian distribution boosts the performance of the proposed approach in practical settings. A theoretical analysis of the BSVT iteration steps describes the conditions under which the proposed method converges and outperforms the SVT-based recovery. Despite the fact that the computational cost per iteration is larger for the new algorithm, the speed of convergence is significantly increased which results in a similar computation time when compared to the SVT algorithm. The numerical experiments presented show that the BSVT algorithm provides robust recovery in realistic scenarios that include noisy observations, mismatched prior knowledge, and non-uniform sampling patterns. In view of this, the new recovery method proposed in this thesis is more suitable to address the challenges posed by the electricity distribution system. It is also worth mentioning that the applicability of the new algorithm is not limited to recovering

missing data in the low voltage electricity grid i.e., BSVT has been used to forecast the traffic levels on the motorway using MIDAS data.

A comprehensive approach to the missing data problem in urban systems is proposed for joint recovery of datasets generated by different subsystems. The novel approach to missing data recovery exploits the correlation that arises as a consequence of the interdependence between the different urban subsystems. Moreover, the joint recovery setting supports more flexibility in the number of observations required from each subsystem. The fundamental limits of the joint recovery of two datasets are studied to provide theoretical guarantees for the joint recovery to require fewer observations compared to the independent case in a matrix completion framework. The description of the recovery regions shows that the joint recovery setting boosts the resilience of the urban system by enabling the recovery of missing data in cases in which it would not be possible in the independent setting.

Furthermore, when the locations of the missing entries are uniformly distributed in each dataset, the probability of recovery is characterized for both the independent and the joint recovery. This result enables the minimization of the sampling complexity for the recovery of two datasets in both the independent and joint recovery setting. Numerical simulations show that the acquisition cost is smaller for the joint recovery in most of the cases compared to the independent case.

In addition, the numerical performance of the BSVT and SVT algorithms is evaluated in the joint setting using synthetic data for different values of intra and cross-correlation. To that end, a new mathematical model that generates correlated datasets is introduced. The performance in exploiting the cross-correlation is evaluated by comparing the numerical results with the fundamental limit. The numerical analysis shows that the BSVT algorithm provides robust recovery in a wider range of sampling regimes in the datasets combined when compared to the SVT-based recovery. Remarkably, the numerical performance of BSVT recovery matches the geometry of the regions described by the fundamental limits which suggests that BSVT is indeed better suited than SVT to exploit the cross-correlation between the combined datasets. In view of this, the BSVT algorithm is a robust alternative to state-of-the-art recovery methods with various applications in the urban setting.

## 6.1 Future work

This section presents several possible extensions to the results presented in this thesis.

1. **Provide probabilistic convergence guarantees for BSVT**. This can be tackled by extending the results in Theorem 1 in a probabilistic framework

using Random Matrix theory. The conditions in Theorem 1 can be relaxed in a probabilistic setting which will provide convergence guarantees for BSVT using a wider range of data matrices.

2. **Optimize the computational effort required by the BSVT algorithm**. In comparison to SVT, the BSVT algorithm has an increased computational cost per iteration due to the Bayesian estimation step incorporated. However, the number of iterations required is smaller for the new algorithm which leads to relatively similar computational cost for both methods. The SVT algorithm exploits the sparsity of the matrix $\mathbf{Y}$ to compute the singular value decomposition faster. In contrast, for the BSVT algortihm the corresponding matrix i.e., $\mathbf{Z}$, is not sparse and new techniques for optimizing the computational effort for the singular value decomposition step are required.

3. **Evaluate the numerical performance of the BSVT algorithm for non-Gaussian low rank data matrices**. The numerical results presented in this thesis pertain to matrices that are low rank (or approximately low rank) and that follow a multivariate Gaussian distribution. In fact, the LMMSE estimation step incorporated into the new algorithm is optimal in this context. However, the applicability of the novel approach is not restricted to this type of matrices. Evaluating the performance of BSVT using non-Gaussian low rank data matrices would provide additional insight into the performance of the proposed algorithm in realistic scenarios.

4. **Extend probability recovery bounds to Markovian sampling models**. In this thesis, lower bounds for the probability of recovery for both the independent and the joint recovery are presented for the case in which the locations of the missing entries from each dataset are sampled uniformly at random. However, this type of sampling has limited applicability in a realistic scenario. In practice, a sensor failure or a downtime in the communication line leads to number of consecutive missing observations. This scenario cannot be modelled using the uniform sampling strategy. In this context, Chapter 4 introduces a non-uniform sampling strategy that is based on a two state Markov chain. Providing recovery guarantees under the Markovian sampling model would provide more insight into the challenges faced by the data acquisition systems in practical settings.

5. **Extend the joint recovery framework to more than two datasets**. In Chapter 5 the joint recovery problem is studied when only two datasets are combined. However, the number of subsystems in an urban setting is significantly larger. That being the case, the joint estimation setting is envisioned

to integrate a larger number subsystems in the recovery framework. The first step would to reproduce the results in Chapter 5 using three datasets with the final objective of generalizing the results for $n$ datasets.

# Bibliography


[1] G. K. Heilig, "World urbanization prospects: The 2011 revision. New York: United Nations, Department of Economic and Social Affairs, Population Division," *Population Estimates and Projections Section*, June 2012.

[2] H. Arbabi and M. Mayfield, "Urban and rural population and energy consumption dynamics in local authorities within England and Wales," *Buildings*, vol. 6, no. 3, p. 34, Aug. 2016.

[3] United Nations Population Division, "World urbanization prospects: the 2014 revision," July 2014.

[4] V. Marchal, R. Dellink, D. van Vuuren, C. Clapp, J. Château, E. Lanzi, B. Magné, and J. van Vliet, "OECD environmental outlook to 2050: The consequences of inaction," Mar. 2012.

[5] H. Ahvenniemi, A. Huovila, I. Pinto-Seppä, and M. Airaksinen, "What are the differences between sustainable and smart cities?," *Cities*, vol. 60, pp. 234–245, Feb. 2017.

[6] J. H. Lee and M. G. Hancock, "Toward a framework for smart cities: A comparison of Seoul, San Francisco and Amsterdam," *Research Paper, Yonsei University and Stanford University*, June 2012.

[7] C. Harrison, B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszczak, and P. Williams, "Foundations for smarter cities," *IBM Journal of Research and Development*, vol. 54, no. 4, pp. 1–16, July 2010.

[8] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, and H. J. Scholl, "Understanding smart cities: An integrative framework," in *Proc. International Conference on System Science*, (Hawaii), pp. 2289–2297, Jan. 2012.

[9] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, Nov. 2013.

[10] J. Qi, J. Wang, and K. Sun, "Efficient estimation of component interactions for cascading failure analysis by EM algorithm," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3153–3161, May 2018.

[11] US-Canada Power System Outage Task Force, *Final report on the August 14, 2003 Blackout in the United States and Canada: causes and recommendations*. Apr. 2004.

[12] Union for the Co-ordination of Transmission of Electricity Investigation Committee, "Final report of the Investigation Committee on the 28 September 2003 Blackout in Italy," *Union for the Co-ordination of Transmission of Electricity*, Apr. 2004.

[13] Federal Energy Regulatory Commission, "Arizona-Southern California outages on September 8, 2011: Causes and recommendations," *Federal Energy Regulatory Commission and The North American Electric Reliability Corporation staff*, Apr. 2012.

[14] L. L. Lai, H. T. Zhang, C. S. Lai, F. Y. Xu, and S. Mishra, "Investigation on July 2012 Indian blackout," in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 1, pp. 92–97, July 2013.

[15] P. Barnaghi, M. Bermudez-Edo, and R. Tönjes, "Challenges for quality of data in smart cities," *Journal of Data and Information Quality*, vol. 6, no. 2-3, p. 6, July 2015.

[16] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks with incomplete information against smart power grids," in *Proc. IEEE Global Communications Conference*, pp. 3153–3158, Dec. 2012.

[17] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Sparse attack construction and state estimation in the smart grid: Centralized and distributed models," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1306–1318, July 2013.

[18] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures," in *Proc. IEEE International Conference on Smart Grid Communications*, pp. 220–225, Oct. 2010.

[19] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, "Recovering missing data via matrix completion in electricity distribution systems," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications*, pp. 1–6, July 2016.

[20] Y. Isozaki, S. Yoshizawa, Y. Fujimoto, H. I. Ishii, I. Ono, T. Onoda, and Y. Hayashi, "Detection of cyber-attacks against voltage control in distribution power grids with PVs," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1824–1835, July 2016.

[21] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1006–1013, Mar. 2016.

[22] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, "Robust recovery of missing data in electricity distribution systems," *IEEE Transactions on Smart Grid*, May 2018.

[23] M. Batty, "Cities as complex systems: Scaling, interaction, networks, dynamics and urban morphologies," Feb. 2008.

[24] P. D. Allison, *Missing data*. Thousand Oaks, CA: Sage publications, 2001.

[25] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 9, no. 10, pp. 617–621, Oct. 1979.

[26] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549–576, Jan. 2009.

[27] Y. Zhou, *Data-driven cyber-physical-social system for knowledge discovery in smart cities*. PhD thesis, University of Surrey, 2018.

[28] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons, 2014.

[29] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, June 2001.

[30] Z. Zhang, "Multiple imputation with multivariate imputation by chained equation (MICE) package," *Annals of translational medicine*, vol. 4, no. 2, Jan. 2016.

[31] X. Yin, D. Levy, C. Willinger, A. Adourian, and M. G. Larson, "Multiple imputation and analysis for high-dimensional incomplete proteomics data," *Statistics in medicine*, vol. 35, no. 8, pp. 1315–1326, Apr. 2016.

[32] P. D. Allison, "Handling missing data by maximum likelihood," in *Proc. SAS global forum*, Apr. 2012.

[33] P. Royston, "Multiple imputation of missing values," *Stata journal*, vol. 4, no. 3, pp. 227–41, Mar. 2004.

[34] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How many imputations are really needed? some practical clarifications of multiple imputation theory," *Prevention science*, vol. 8, no. 3, pp. 206–213, Sept. 2007.

[35] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, Apr. 2009.

[36] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[37] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2057–2078, July 2010.

[38] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[39] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

[40] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan, and F. Bu, "A survey on state estimation techniques and challenges in smart distribution systems," *arXiv preprint arXiv:1809.00057*, Aug. 2018.

[41] V. Kekatos, Y. Zhang, and G. B. Giannakis, "Electricity market forecasting via low-rank multi-kernel learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 6, pp. 1182–1193, Dec. 2014.

[42] Z. Liu, Y. Yan, J. Yang, and M. Hauskrecht, "Missing value estimation for hierarchical time series: A study of hierarchical web traffic," in *Proc. IEEE International Conference on Data Mining*, pp. 895–900, Nov. 2015.

[43] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, Aug. 2004.

[44] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, Jan. 1895.

[45] H. Kurasawa, H. Sato, A. Yamamoto, H. Kawasaki, M. Nakamura, Y. Fujii, and H. Matsumura, "Missing sensor value estimation method for participatory sensing environment," in *Proc. IEEE International Conference on Pervasive Computing and Communications*, pp. 103–111, Mar. 2014.

[46] A. Chistov and D. Grigoriev, "Complexity of quantifier elimination in the theory of algebraically closed fields," in *Proc. Symposium on Mathematical Foundation of Computer Science*, (Praha, Czechoslovakia), pp. 17–31, Sept. 1984.

[47] C. Beck and R. D'Andrea, "Computational study and comparisons of LFT reducibility methods," in *Proc. American Control Conference*, vol. 2, pp. 1013–1017, June 1998.

[48] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. American Control Conference*, vol. 6, pp. 4734–4739, June 2001.

[49] M. Fazel, *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[50] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, June 2010.

[51] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, pp. 3413–3430, Dec. 2011.

[52] A. Rohde and A. Tsybakov, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, Apr. 2011.

[53] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, Sept. 2011.

[54] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, May 2012.

[55] R. Foygel and N. Srebro, "Concentration-based guarantees for low-rank matrix reconstruction," in *Proc. Annual Conference on Learning Theory*, pp. 315–340, Dec. 2011.

[56] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, Feb. 2014.

[57] T. T. Cai and W. Zhou, "Matrix completion via max-norm constrained optimization," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1493–1525, May 2016.

[58] O. Klopp, K. Lounici, and A. Tsybakov, "Robust matrix completion," *Probability Theory and Related Fields*, vol. 169, no. 1-2, pp. 523–564, Oct. 2017.

[59] E. Riegler, D. Stotz, and H. Bölcskei, "Information-theoretic limits of matrix completion," in *Proc. IEEE International Symposium on Information Theory*, pp. 1836–1840, June 2015.

[60] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[61] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Transactions on Sensor Networks*, vol. 2, no. 2, pp. 188–220, May 2006.

[62] A. Singer, "A remark on global positioning from local distances," *Proceedings of the National Academy of Sciences*, vol. 105, no. 28, pp. 9507–9511, July 2008.

[63] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 1621–1641, Feb. 2010.

[64] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, Nov. 2009.

[65] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[66] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *Proc. American Control Conference*, vol. 4, pp. 3273–3278, June 2004.

[67] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[68] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[69] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: application to SFM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1051–1063, Aug. 2004.

[70] J. Abernethy, F. Bach, T. Evgeniou, and J. Vert, "Low-rank matrix factorization with attributes," *arXiv preprint cs/0611124*, Nov. 2006.

[71] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Advances in Neural Information Processing Systems*, pp. 41–48, Sept. 2007.

[72] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. International Conference on Machine learning*, pp. 17–24, June 2007.

[73] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[74] D. Kressner, M. Steinlechner, and B. Vandereycken, "Low-rank tensor completion by Riemannian optimization," *BIT Numerical Mathematics*, vol. 54, no. 2, pp. 447–468, June 2014.

[75] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 148–172, Feb. 2013.

[76] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3842–3849, June 2014.

[77] M. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3367–3381, June 2017.

[78] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. International Conference on Knowledge Discovery and Data Mining*, pp. 650–658, Aug. 2008.

[79] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang, "Consistent collective matrix completion under joint low rank structure," in *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 306–314, Feb. 2015.

[80] K. C. Toh, M. J. Todd, and R. H. Tutuncu, "SDPT3 - A MATLAB software package for semidefinite programming, version 1.3," *Optimization Methods & Software*, vol. 11, no. 1-4, pp. 545–581, Jan. 1999.

[81] J. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods & Software*, vol. 11, no. 1-4, pp. 625–653, Jan. 1999.

[82] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, Feb. 2018.

[83] K. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of optimization*, vol. 6, no. 615-640, p. 15, Sept. 2010.

[84] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[85] N. Rao, P. Shah, and S. Wright, "Forward-backward greedy algorithms for atomic norm regularization," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5798–5811, Nov. 2015.

[86] N. Boyd, G. Schiebinger, and B. Recht, "The alternating descent conditional gradient method for sparse inverse problems," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 616–639, Apr. 2017.

[87] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, Sept. 2010.

[88] M. Jaggi, "Revisiting Frank-Wolfe: projection-free sparse convex optimization," in *Proc. International Conference on Machine Learning*, pp. I–427, June 2013.

[89] R. M. Freund, P. Grigas, and R. Mazumder, "An extended Frank-Wolfe method with "in-face" directions, and its application to low-rank matrix completion," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 319–346, Mar. 2017.

[90] Z. Allen-Zhu, E. Hazan, W. Hu, and Y. Li, "Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls," in *Proc. Advances in Neural Information Processing Systems*, pp. 6191–6200, Dec. 2017.

[91] S. Burer and R. D. C. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Mathematical Programming*, vol. 103, no. 3, pp. 427–444, July 2005.

[92] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, Sept. 2015.

[93] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," *arXiv preprint arXiv:1711.10467*, Nov. 2017.

[94] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. Annual ACM Symposium on Theory of Computing*, pp. 665–674, June 2013.

[95] M. Hardt, "Understanding alternating minimization for matrix completion," in *Proc. IEEE Annual Symposium on Foundations of Computer Science*, pp. 651–660, Oct. 2014.

[96] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proc. Advances in Neural Information Processing Systems*, pp. 937–945, Dec. 2010.

[97] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Proc. Conference on Learning Theory*, pp. 1007–1034, June 2015.

[98] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Proc. Advances in Neural Information Processing Systems*, pp. 1107–1115, Dec. 2014.

[99] Electricity North West Limited, "Low voltage network solutions," [Online]. Available: https://www.enwl.co.uk/innovation/smaller-projects/low-carbon-networks-fund/low-voltage-network-solutions/. Accessed: 22 Sept. 2018.

[100] G. A. F. Seber, *A matrix handbook for statisticians*. Hoboken, NJ: Wiley, 2007.

[101] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, July 1948.

[102] T. Cover and J. Thomas, *Elements of information theory*. Hoboken, NJ: John Wiley & Sons, 2012.

[103] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Transactions on Information Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.

[104] I. Esnaola, A. M. Tulino, and H. V. Poor, "Mismatched MMSE estimation of multivariate Gaussian sources," in *Proc. IEEE International Symposium on Information Theory*, pp. 716–720, July 2012.

[105] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, Aug. 2010.

[106] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *The Annals of Statistics*, vol. 42, no. 6, pp. 2413–2440, Dec. 2014.

[107] X. Jia, X. Feng, and W. Wang, "Rank constrained nuclear norm minimization with application to image denoising," *Signal Processing*, vol. 129, no. 1, pp. 1–11, Dec. 2016.

[108] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[109] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko, "Unbiased risk estimates for singular value thresholding and spectral estimators," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4643–4657, Oct. 2013.

[110] P. R. Halmos, *Measure theory*. New York, NY: Springer Science & Business Media, 2013.

[111] E. Hewitt and K. Stromberg, *Real and abstract analysis: a modern treatment of the theory of functions of a real variable*. Berlin, Germany: Springer-Verlag, 2013.

[112] H. L. Royden and P. Fitzpatrick, *Real analysis*. New York, NY: Pearson, 2018.

[113] Y. H. Wang, "On the number of successes in independent trials," *Statistica Sinica*, pp. 295–312, July 1993.

[114] S. Tucker, I. Summersgill, J. Fletcher, and D. Mustard, "Evaluating the benefits of MIDAS automatic queue protection," *Traffic engineering & control*, vol. 47, no. 9, Oct. 2006.

[115] A. L. Clements, W. G. Griswold, A. Rs, J. E. Johnston, M. M. Herting, J. Thorson, A. Collier-Oxandale, and M. Hannigan, "Low-cost air quality monitoring tools: from research to practice (a workshop summary)," *Sensors*, vol. 17, no. 11, p. 2478, Oct. 2017.

[116] S. Jorgensen, R. H. Chen, M. B. Milam, and M. Pavone, "The team surviving orienteers problem: routing teams of robots in uncertain environments with survival constraints," *Autonomous Robots*, vol. 42, no. 4, pp. 927–952, Dec. 2017.

[117] F. Zhang, *The Schur complement and its applications*. New York, NY: Springer Science & Business Media, 2006.

# Appendix A

# Proofs

## Proofs for Chapter 4

In the following we present two auxiliary lemmas that are instrumental in the proof of Theorem 1. To that end, let us recall the iterations steps of the SVT algorithm:

$$\begin{cases} \mathbf{X}_s^{(k)} = D_\tau(\mathbf{Y}_s^{(k-1)}), \\ \mathbf{Y}_s^{(k)} = \mathbf{Y}_s^{(k-1)} + \delta_s P_\Omega(\mathbf{M} - \mathbf{X}_s^{(k)}), \end{cases} \tag{A.1}$$

where the initialization point is chosen as $\mathbf{Y}_s^{(0)} = \mathbf{0}$, $\delta_s$ is the step size that obeys $0 \leq \delta_s \leq 2$, and the soft-thresholding operator, $D_\tau$, is defined in the following.

For a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ of rank $r$ with singular value decomposition given by

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad \mathbf{S} = \text{diag}(\{\sigma_i(\mathbf{X})\}_{1 \leq i \leq r}), \tag{A.2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices with orthogonal columns of size $M \times r$ and $N \times r$, respectively, and $\sigma_i(\cdot)$ denotes the $i$th singular value of the matrix, the soft-thresholding operator is defined as

$$D_\tau(\mathbf{X}) := \mathbf{U} D_\tau(\mathbf{S}) \mathbf{V}^T, \quad D_\tau(\mathbf{S}) = \text{diag}(\{(\sigma_i(\mathbf{X}) - \tau)_+\}), \tag{A.3}$$

where $t_+ = \max\{0, t\}$.

For the proposed approach, i.e., BSVT, the iterations of the algorithm are

$$\begin{cases} \mathbf{X}_b^{(k)} = D_\tau(\mathbf{Z}^{(k-1)}), \\ \mathbf{Y}_b^{(k)} = \mathbf{Y}_b^{(k-1)} + \delta_b P_\Omega(\mathbf{M} - \mathbf{X}_b^{(k)}), \\ \mathbf{Z}^{(k)} = \mathbf{Y}_b^{(k)} + \mathbf{L}^{(k)}, \end{cases} \tag{A.4}$$

where the initialization point is chosen as $\mathbf{Y}_b^{(0)} = \mathbf{0}$ and $\mathbf{Z}^{(0)} = \mathbf{0}$, $\delta_b$ is the step size, the soft-thresholding operator, $D_\tau$, is defined in (A.3), and $\mathbf{L}^{(k)}$ is defined in (4.2).

Using the singular value inequalities in 6.72b(iii) in [100] for the matrix $\mathbf{Z}^{(k)}$ gives

$$\sigma_i(\mathbf{Z}^{(k)}) \leq \sigma_i(\mathbf{Y}_b^{(k)}) + \sigma_{\max}(\mathbf{L}^{(k)}), \tag{A.5}$$

and

$$\sigma_i(\mathbf{Y}_b^{(k)}) + \sigma_{\min}(\mathbf{L}^{(k)}) \leq \sigma_i(\mathbf{Z}^{(k)}), \tag{A.6}$$

which leads to

$$\sigma_i(\mathbf{Y}_b^{(k)}) \leq \sigma_i(\mathbf{Z}^{(k)}), \tag{A.7}$$

where $\sigma_i(\cdot)$ denotes the $i$th singular value, $\sigma_{\max}(\cdot)$ denotes the maximum singular value and $\sigma_{\min}(\cdot)$ denotes the minimum singular value.

**Lemma 6.** *Given the BSVT iteration steps defined in (A.4), the following singular value inequality holds*

$$\sigma_i(\mathbf{X}_b^{(k)}) \leq \sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)})) + \sigma_{\max}(\mathbf{L}^{(k-1)}), \tag{A.8}$$

*for every iteration $k$.*

*Proof.* Given the definition of the matrix $\mathbf{X}_b^{(k)}$ in (A.4), the singular values of $\mathbf{X}_b^{(k)}$ are given by

$$\sigma_i(\mathbf{X}_b^{(k)}) = (\sigma_i(\mathbf{Z}^{(k-1)}) - \tau)_+. \tag{A.9}$$

Using the inequality from $(A.5)$, leads to

$$(\sigma_i(\mathbf{Z}^{(k-1)}) - \tau)_+ \leq (\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\max}(\mathbf{L}^{(k-1)}) - \tau)_+, \tag{A.10}$$

which can be further extended as

$$(\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\max}(\mathbf{L}^{(k-1)}) - \tau)_+ \leq (\sigma_i(\mathbf{Y}_b^{(k-1)}) - \tau)_+ + \sigma_{\max}(\mathbf{L}^{(k-1)}). \tag{A.11}$$

Combining (A.10) and (A.11) gives

$$(\sigma_i(\mathbf{Z}^{(k-1)}) - \tau)_+ \leq (\sigma_i(\mathbf{Y}_b^{(k-1)}) - \tau)_+ + \sigma_{\max}(\mathbf{L}^{(k-1)}), \tag{A.12}$$

which concludes the proof. □

**Lemma 7.** *Given the BSVT iteration steps defined in (A.4), there exists a step size $\delta_b$ such that*

$$\sigma_i(\mathbf{Y}_b^{(k)}) \leq \sigma_i(\mathbf{Y}_s^{(k)}), \tag{A.13}$$

*for every iteration $k$.*

*Proof.* Note that for the first iteration, i.e. $k = 1$, we have that

$$\mathbf{Y}_s^{(1)} = \delta_s P_\Omega(\mathbf{M}), \tag{A.14}$$

and

$$\mathbf{Y}_b^{(1)} = \delta_b P_\Omega(\mathbf{M}). \tag{A.15}$$

It can be shown that $\sigma_i(\mathbf{Y}_b^{(1)}) \leq \sigma_i(\mathbf{Y}_s^{(1)})$ when $\delta_b \leq \delta_s$. Therefore, there exists a value of $\delta_b$ such that

$$\sigma_i(\mathbf{Y}_b^{(1)}) \leq \sigma_i(\mathbf{Y}_s^{(1)}). \tag{A.16}$$

Suppose that for iteration $k$ the following inequality is satisfied

$$\sigma_i(\mathbf{Y}_b^{(k)}) \leq \sigma_i(\mathbf{Y}_s^{(k)}). \tag{A.17}$$

We proceed by induction and show that

$$\sigma_i(\mathbf{Y}_b^{(k+1)}) \leq \sigma_i(\mathbf{Y}_s^{(k+1)}). \tag{A.18}$$

Given the iterations of the SVT algorithm in (A.1), the matrix $\mathbf{Y}_s$ at iteration $(k+1)$ is given by

$$\mathbf{Y}_s^{(k+1)} = \mathbf{Y}_s^{(k)} + \delta_s P_\Omega(\mathbf{M} - \mathbf{X}_s^{(k+1)}). \tag{A.19}$$

Using the singular value inequalities in 6.72b(iii) in [100] for the matrix $\mathbf{Y}_s^{(k+1)}$ yields

$$\sigma_i(\mathbf{Y}_s^{(k+1)}) \geq \sigma_i(\mathbf{Y}_s^{(k)}) + \sigma_{\min}(\delta_s P_\Omega(\mathbf{M} - \mathbf{X}_s^{(k+1)})). \tag{A.20}$$

For the BSVT algorithm, $\mathbf{Y}_b$ at iteration $k+1$ is defined as

$$\mathbf{Y}_b^{(k+1)} = \mathbf{Y}_b^{(k)} + \delta_b P_\Omega(\mathbf{M} - \mathbf{X}_b^{(k+1)}), \tag{A.21}$$

and using the singular value inequalities in 6.72b(iii) in [100] for the matrix $\mathbf{Y}_b^{(k+1)}$ gives

$$\sigma_i(\mathbf{Y}_b^{(k+1)}) \leq \sigma_i(\mathbf{Y}_b^{(k)}) + \sigma_{\max}(\delta_b P_\Omega(\mathbf{M} - \mathbf{X}_b^{(k+1)})). \tag{A.22}$$

Combining (A.20) and (A.22) with (A.17) results in the claim (A.18) being equivalent to

$$\sigma_{\max}(\delta_b P_\Omega(\mathbf{M} - \mathbf{X}_b^{(k+1)})) \leq \sigma_{\min}(\delta_s P_\Omega(\mathbf{M} - \mathbf{X}_s^{(k+1)})), \tag{A.23}$$

which is equivalent to

$$\frac{\delta_b}{\delta_s}\sigma_{\max}(P_\Omega(\mathbf{M} - \mathbf{X}_b^{(k+1)})) \leq \sigma_{\min}(P_\Omega(\mathbf{M} - \mathbf{X}_s^{(k+1)})). \tag{A.24}$$

Since $\sigma_{\min}(P_\Omega(\mathbf{M}-\mathbf{X}_s^{(k+1)}) \geq 0$, there exists a value for $\delta_b$ for which the inequality is satisfied. To guarantee that the inequality is satisfied at each iteration $k$, the value of $\delta_b$ is the minimum value that satisfies the inequality in (A.24) for $k \in \{1, 2, ..., k_{\max}\}$. □

**Theorem 1.** *Let $\mathbf{X}_b^{(k)}$ be the estimate produced by the BSVT algorithm at iteration $k$ and $\mathbf{X}_s^{(k)}$ be the estimate produced by the SVT algorithm at iteration $k$ with $\mathbf{L}^{(k)}$ denoting the LMMSE estimation output at iteration $k$ as defined in (4.2). If the following holds:*

$$tr\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \geq tr\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big), \tag{A.25}$$

*and*

$$\lim_{k\to\infty} \sigma_{\min}(\mathbf{L}^{(k)}) = 0, \tag{A.26}$$

*then*

$$\|\mathbf{X}_b^{(k)} - \mathbf{M}\|_F^2 \leq \|\mathbf{X}_s^{(k)} - \mathbf{M}\|_F^2. \tag{A.27}$$

*Proof.* The inequality in (A.27) can be expanded as

$$\mathrm{tr}((\mathbf{X}_b^{(k)} - \mathbf{M})(\mathbf{X}_b^{(k)} - \mathbf{M})^T) \leq \mathrm{tr}((\mathbf{X}_s^{(k)} - \mathbf{M})(\mathbf{X}_s^{(k)} - \mathbf{M})^T), \tag{A.28}$$

which is equivalent to

$$\|\mathbf{X}_b^{(k)}\|_F^2 + \|\mathbf{M}\|_F^2 - 2\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \|\mathbf{X}_s^{(k)}\|_F^2 + \|\mathbf{M}\|_F^2 - 2\mathrm{tr}\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big), \tag{A.29}$$

where $\|\mathbf{M}\|_F^2$ can be subtracted from both sides resulting in

$$\|\mathbf{X}_b^{(k)}\|_F^2 - 2\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \|\mathbf{X}_s^{(k)}\|_F^2 - 2\mathrm{tr}\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big). \tag{A.30}$$

Note that

$$\|\mathbf{X}_b^{(k)}\|_F^2 = \sum_{i=1}^{n} \sigma_i^2(\mathbf{X}_b^{(k)}), \tag{A.31}$$

where

$$\sigma_i(\mathbf{X}_b^{(k)}) = (\sigma_i(\mathbf{Z}^{(k-1)}) - \tau)_+, \tag{A.32}$$

and $n$ denotes the number of singular values of the matrix $\mathbf{X}_b^{(k)}$. Using 6.72b(iii) in [100] for the matrix $\mathbf{X}_b^{(k)}$ gives

$$\sigma_i(\mathbf{X}_b^{(k)}) \geq (\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)_+, \tag{A.33}$$

where

$$(\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)_+ = \begin{cases} (\sigma_i(\mathbf{Y}_b^{(k-1)}) - \tau)_+ + \sigma_{\min}(\mathbf{L}^{(k-1)}), & \text{for } \sigma_i(\mathbf{Y}_b^{(k-1)}) \geq \tau \\ \sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau, & \text{for } \begin{cases} \sigma_i(\mathbf{Y}_b^{(k-1)}) & < \tau \\ \sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) & > \tau \end{cases} \\ 0, & \text{otherwise.} \end{cases} \tag{A.34}$$

Without loss of generality, assume that the singular values of the matrix $\mathbf{Y}_b^{(k-1)}$ are ordered such that $\sigma_i(\mathbf{Y}_b^{(k-1)}) \geq \sigma_j(\mathbf{Y}_b^{(k-1)})$ for $i \leq j$ (i.e., descending order) and define $n_1$ and $n_2$ such that $\sigma_i(\mathbf{Y}_b^{(k-1)}) \geq \tau$ for $i = \overline{1, n_1}$ and $\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) > \tau$ for $i = \overline{n_1+1, n_2}$. Therefore

$$\|\mathbf{X}_b^{(k)}\|_F^2 = \sum_{i=1}^{n_1} \sigma_i^2(\mathbf{X}_b^{(k)}) + \sum_{i=n_1+1}^{n_2} \sigma_i^2(\mathbf{X}_b^{(k)}) + \sum_{i=n_2+1}^{n} \sigma_i^2(\mathbf{X}_b^{(k)}), \tag{A.35}$$

which leads to

$$\begin{aligned} \|\mathbf{X}_b^{(k)}\|_F^2 \geq & \sum_{i=1}^{n_1} \Big( (\sigma_i(\mathbf{Y}_b^{(k-1)}) - \tau)_+ + \sigma_{\min}(\mathbf{L}^{(k-1)}) \Big)^2 + \\ & \sum_{i=n_1+1}^{n_2} \Big( \sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau \Big)^2, \end{aligned} \tag{A.36}$$

that is equivalent with

$$\begin{aligned} \|\mathbf{X}_b^{(k)}\|_F^2 \geq & \|D_\tau(\mathbf{Y}_b^{(k-1)})\|_F^2 + 2\sum_{i=1}^{n_1} \sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{\min}(\mathbf{L}^{(k-1)}) + n_1\sigma_{\min}^2(\mathbf{L}^{(k-1)}) + \\ & \sum_{i=n_1+1}^{n_2} (\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)^2. \end{aligned} \tag{A.37}$$

Combining (A.30) with (A.37) gives

$$\begin{aligned} \|D_\tau(\mathbf{Y}_b^{(k-1)})\|_F^2 + n_1\sigma_{\min}^2(\mathbf{L}^{(k-1)}) + 2\sum_{i=1}^{n_1} \sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{\min}(\mathbf{L}^{(k-1)}) + \\ \sum_{i=n_1+1}^{n_2} (\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)^2 \leq \|D_\tau(\mathbf{Y}_s^{(k-1)})\|_F^2 + \\ 2\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) - 2\mathrm{tr}\Big(D_\tau(\mathbf{Y}_s^{(k-1)})\mathbf{M}^T\Big). \end{aligned} \tag{A.38}$$

Using Lemma 7, the following inequality holds

$$\|D_\tau(\mathbf{Y}_b^{(k-1)})\|_F^2 \leq \|D_\tau(\mathbf{Y}_s^{(k-1)})\|_F^2, \tag{A.39}$$

which implies that the inequality in (A.38) is satisfied if

$$\begin{aligned} &n_1\sigma_{\min}^2(\mathbf{L}^{(k-1)}) + 2\sum_{i=1}^{n_1}\sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{\min}(\mathbf{L}^{(k-1)}) \\ &+ \sum_{i=n_1+1}^{n_2}(\sigma_i(\mathbf{Y}_b^{(k-1)}) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)^2 \leq 2\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) - 2\mathrm{tr}\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big). \end{aligned} \tag{A.40}$$

Let $\lim_{k\to\infty}\sigma_{\min}(\mathbf{L}^{(k)}) = 0$ then

$$n_1\sigma_{\min}^2(\mathbf{L}^{(k-1)}) \to 0. \tag{A.41}$$

Moreover, it also holds that

$$2\sum_{i=1}^{n_1}\sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{\min}(\mathbf{L}^{(k-1)}) \to 0, \tag{A.42}$$

and since $(\sigma_i(\mathbf{Y}_b) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau) < \sigma_{\min}(\mathbf{L}^{(k-1)})$ for $n_1 < i \leq n_2$, then

$$\sum_{i=n_1+1}^{n_2}(\sigma_i(\mathbf{Y}_b) + \sigma_{\min}(\mathbf{L}^{(k-1)}) - \tau)^2 \to 0. \tag{A.43}$$

Consequently, the inequality in (A.40) becomes

$$0 \leq 2\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) - 2\mathrm{tr}\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big), \tag{A.44}$$

which concludes the proof. □

**Corollary 1.1.** *Let $k_s$ be the last iteration of the SVT algorithm and denote by $\mathbf{X}_s^{(k_s)}$ the output of the SVT algorithm. For the case in which $\mathbf{X}_b^{(k)}\mathbf{M}^T$ and $\mathbf{X}_s^{(k)}\mathbf{M}^T$ are Hermitian, the condition (A.25) is equivalent to*

$$0 \leq 2\sum_{i=1}^{r_s}\sigma_i(\mathbf{X}_s^{(k_s)})\Big(\sigma_{\max}(\mathbf{L}^{(k_s)}) - \sigma_i(\mathbf{X}_s^{(k_s)})\Big) + 2\Big(\|\mathbf{X}_s^{(k_s)}\|_* + \sigma_{\max}(\mathbf{L}^{(k_s)})\Big)\varepsilon, \tag{A.45}$$

*where $r_s$ denotes the rank of the matrix $\mathbf{X}_s^{(k_s)}$, $\sigma_i(\mathbf{X}_s^{(k_s)})$ denotes the $i$th singular value of the matrix $\mathbf{X}_s^{(k_s)}$ and $\sigma_{\max}(\mathbf{L}^{(k_s)})$ denotes the maximum singular value of the matrix $\mathbf{L}^{(k_s)}$.*

*Proof.* When the matrix $\mathbf{X}_b^{(k)}\mathbf{M}^T$ is Hermitian, the trace term in (A.25) is given by (6.57 in [100])

$$\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) = \sum_{i=1}^{n}\sigma_i(\mathbf{X}_b^{(k)})\sigma_{p(i)}(\mathbf{M}^T), \tag{A.46}$$

where $p(i)$ is the $i$th element of a permutation of $\{1, 2, ..., n\}$ and $n$ denotes the number of singular values.

Using Lemma 6, (A.46) becomes

$$\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \sum_{i=1}^{n}\Big(\sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)})) + \sigma_{\max}(\mathbf{L}^{(k-1)})\Big)\sigma_{p(i)}(\mathbf{M}^T), \tag{A.47}$$

which leads to

$$\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \sum_{i=1}^{n}\sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{p(i)}(\mathbf{M}^T) + \sigma_{\max}(\mathbf{L}^{(k-1)})\sum_{i=1}^{n}\sigma_{p(i)}(\mathbf{M}^T). \tag{A.48}$$

This is equivalent to

$$\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \sum_{i=1}^{n}\sigma_i(D_\tau(\mathbf{Y}_b^{(k-1)}))\sigma_{p(i)}(\mathbf{M}) + \|\mathbf{M}\|_*\sigma_{\max}(\mathbf{L}^{(k-1)}). \tag{A.49}$$

Combining (A.49) with Lemma 7 leads to

$$\mathrm{tr}\Big(\mathbf{X}_b^{(k)}\mathbf{M}^T\Big) \leq \sum_{i=1}^{n}\sigma_i(D_\tau(\mathbf{Y}_s^{(k-1)}))\sigma_{p(i)}(\mathbf{M}) + \|\mathbf{M}\|_*\sigma_{\max}(\mathbf{L}^{(k-1)}). \tag{A.50}$$

For the second trace term in (A.25), when $\mathbf{X}_s^{(k)}\mathbf{M}^T$ is Hermitian, the same result in 6.57 in [100] gives

$$\mathrm{tr}\Big(\mathbf{X}_s^{(k)}\mathbf{M}^T\Big) = \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{(k)})\sigma_{q(i)}(\mathbf{M}^T), \tag{A.51}$$

where $q(i)$ is the $i$th element of a permutation of $\{1, 2, ..., n\}$.

Combining (A.25), (A.50) and (A.51) gives

$$0 \leq \|\mathbf{M}\|_*\sigma_{\max}(\mathbf{L}^{(k-1)}) + \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{(k)})\sigma_{p(i)}(\mathbf{M}) - \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{(k)})\sigma_{q(i)}(\mathbf{M}). \tag{A.52}$$

Therefore (A.52) is equivalent to

$$0 \leq \|\mathbf{M}\|_*\sigma_{\max}(\mathbf{L}^{(k-1)}) + \beta, \tag{A.53}$$

with

$$\beta = \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{k})\sigma_{p(i)}(\mathbf{M}) - \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{k})\sigma_{q(i)}(\mathbf{M}), \tag{A.54}$$

where $p(i)$ is the $i$th element of a permutation of $\{1, 2, ..., n\}$ and $q(i)$ is the $i$th element of a permutation of $\{1, 2, ..., n\}$.

**Lemma 8.** *Given the definition of $\beta$ in (A.54), the minimum value of $\beta$ is given by $\min_{p,q}\beta = \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{(k)})\sigma_{n-i+1}(\mathbf{M}) - \sum_{i=1}^{n}\sigma_i(\mathbf{X}_s^{(k)})\sigma_i(\mathbf{M})$, when the singular values of both matrices are ordered in descending order.*

*Proof.* We proceed to prove the result by contradiction. Since $\beta$ is defined as the subtraction of two positive terms, the minimum is given by

$$\min_{p,q} \beta = \min_{p} \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{p(i)}(\mathbf{M}) - \max_{q} \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{q(i)}(\mathbf{M}). \tag{A.55}$$

Without loss of generality, consider that the singular values of $\mathbf{X}_s^{(k)}$ are ordered in descending order. Denote by

$$\alpha = \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{p(i)}(\mathbf{M}), \tag{A.56}$$

where $\sigma_i(\mathbf{X}_s^{(k)})$ are ordered in descending order and $\sigma_{p(i)}(\mathbf{M})$ are ordered in ascending order.
Assume that there is a permutation $p\prime$ in which positions $i$ and $j$ are switched such that

$$\alpha\prime = \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{p\prime(i)}(\mathbf{M}) \tag{A.57}$$

satisfies $\alpha\prime \leq \alpha$. This is equivalent to

$$\sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{p\prime(i)}(\mathbf{M}) \leq \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{p(i)}(\mathbf{M}). \tag{A.58}$$

After some algebraic manipulations, the inequality boils down to

$$\sigma_i(\mathbf{X}_s^{(k)})\Big(\sigma_{p(j)}(\mathbf{M}) - \sigma_{p(i)}(\mathbf{M})\Big) \leq \sigma_j(\mathbf{X}_s^{(k)})\Big(\sigma_{p(j)}(\mathbf{M}) - \sigma_{p(i)}(\mathbf{M})\Big). \tag{A.59}$$

But

$$\sigma_{p(j)}(\mathbf{M}) \geq \sigma_{p(i)}(\mathbf{M}), \tag{A.60}$$

and

$$\sigma_i(\mathbf{X}_s^{(k)}) \geq \sigma_j(\mathbf{X}_s^{(k)}), \tag{A.61}$$

which contradicts the assumption that $\alpha\prime \leq \alpha$.
Similarly to the minimum,

$$\max_{q} \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{q(i)}(\mathbf{M}) \tag{A.62}$$

is obtained when both $\sigma_i(\mathbf{X}_s^{(k)})$ and $\sigma_{q(i)}(\mathbf{M})$ are ordered in descending order. Therefore

$$\min_{p,q} \beta = \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{n-i+1}(\mathbf{M}) - \sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_i(\mathbf{M}), \tag{A.63}$$

when $\sigma_i(\mathbf{X}_s^{(k)})$ and $\sigma_i(\mathbf{M})$ are ordered in descending order. □

Using Lemma 8, (A.53) becomes

$$0 \leq 2\|\mathbf{M}\|_* \sigma_{\max}(\mathbf{L}^{(k-1)}) + 2\sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_{n-i+1}(\mathbf{M}) - 2\sum_{i=1}^{n} \sigma_i(\mathbf{X}_s^{(k)})\sigma_i(\mathbf{M}), \quad \text{(A.64)}$$

where the singular values of $\mathbf{X}_s^{(k)}$ and $\mathbf{M}$ are ordered in descending order.

Moreover, when the SVT recovery is successful the following equality is satisfied

$$\sigma_i(\mathbf{M}) = \sigma_i(\mathbf{X}_s) + \varepsilon_i, \quad \text{(A.65)}$$

where $\varepsilon_i$ is a small constant. This is equivalent to

$$\|\mathbf{M}\|_* = \|\mathbf{X}_s\|_* + \varepsilon, \quad \text{(A.66)}$$

where $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$. Combining (A.65) and (A.66) with (A.64) for $k = k_s$ that is last iteration of the SVT algorithm leads to

$$0 \leq \sigma_{\max}(\mathbf{L}^{k_s})(\|\mathbf{X}_s^{k_s}\|_* + \varepsilon) + \sum_{i=1}^{\min\{M,N\}} \sigma_i(\mathbf{X}_s^{k_s})(\sigma_{\min\{M,N\}-i+1}(\mathbf{X}_s^{k_s}) - \sigma_i(\mathbf{X}_s^{k_s}) + \varepsilon_{\min\{M,N\}-i+1} + \varepsilon_i), \quad \text{(A.67)}$$

where the singular values of $\mathbf{X}_s^{(k)}$ and $\mathbf{M}$ are ordered in descending order. □

# Proofs for Chapter 5

**Lemma 1.** *Given two low rank matrices* $\mathbf{M}_1$, $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$ *and the combined matrix* $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2M\times N}$, *the rank of* $\mathbf{M}$ *is bounded by*

$$\max\{\text{rank}(\mathbf{M}_1), \text{rank}(\mathbf{M}_2)\} \leq \text{rank}(\mathbf{M}) \leq \text{rank}(\mathbf{M}_1) + \text{rank}(\mathbf{M}_2). \tag{A.68}$$

*Proof.* The rank of the matrix $\mathbf{M}$ is defined as (3.23(a) in [100])

$$\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{M}_1) + \text{rank}(\mathbf{M}_2) - d, \tag{A.69}$$

where $d$ is the intersection of the row subspaces of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ given by

$$d = \dim[\mathcal{C}(\mathbf{M}_1^T) \cap \mathcal{C}(\mathbf{M}_2^T)]. \tag{A.70}$$

Since, the intersection the two subspaces obeys

$$\dim[\mathcal{C}(\mathbf{M}_1^T) \cap \mathcal{C}(\mathbf{M}_2^T)] \geq 0, \tag{A.71}$$

the rank of the matrix $\mathbf{M}$ satisfies

$$\text{rank}(\mathbf{M}) \leq \text{rank}(\mathbf{M}_1) + \text{rank}(\mathbf{M}_2). \tag{A.72}$$

Moreover, based on (A.69), the rank can be lower bounded by

$$\text{rank}(\mathbf{M}) \geq \text{rank}(\mathbf{M}_1) + \text{rank}(\mathbf{M}_2) - \max\{d\}, \tag{A.73}$$

where $\max\{d\}$ denotes the maximum possible value for the intersection of the row subspaces of $\mathbf{M}_1$ and $\mathbf{M}_2$. Note that

$$\dim[\mathcal{C}(\mathbf{M}_1^T)] = r_1, \tag{A.74}$$

and

$$\dim[\mathcal{C}(\mathbf{M}_2^T)] = r_2. \tag{A.75}$$

Therefore, the intersection of the two subspaces is upper bounded by

$$\dim[\mathcal{C}(\mathbf{M}_1^T) \cap \mathcal{C}(\mathbf{M}_2^T)] \leq \min\{r_1, r_2\}. \tag{A.76}$$

Consequently, the rank of $\mathbf{M}$ is lower bounded by

$$\text{rank}(\mathbf{M}) \geq r_1 + r_2 - \min\{r_1, r_2\}, \tag{A.77}$$

which is equivalent to

$$\text{rank}(\mathbf{M}) \geq \max\{r_1, r_2\}. \tag{A.78}$$

Combining (A.72) and (A.78) gives

$$\max\{r_1, r_2\} \leq \text{rank}(\mathbf{M}) \leq r_1 + r_2. \tag{A.79}$$

□

**Lemma 2.** *(Necessary condition for beneficial joint recovery) Let* $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M \times N}$ *with rank* $r_1$ *and* $r_2$ *respectively, there exists a value of* $r$ *for which the joint recovery is beneficial if*

$$M(\min\{r_1, r_2\} - \max\{r_1, r_2\}) > \min\{r_1, r_2\}(\min\{r_1, r_2\} - N). \tag{A.80}$$

*Proof.* The inequality in (5.19) defines the beneficial joint recovery cases. This is equivalent to

$$(2M + N - r)r < (M + N - r_1)r_1 + (M + N - r_2)r_2. \tag{A.81}$$

Moving all the terms on the left side leads to

$$-r^2 + (2M + N)r - (M + N)(r_1 + r_2) + r_1^2 + r_2^2 < 0. \tag{A.82}$$

The term on the left side is a quadratic with the variable $r$ given by

$$-r^2 + (2M + N)r - (M + N)(r_1 + r_2) + r_1^2 + r_2^2 = 0. \tag{A.83}$$

The sign of the function depends on $\Delta$ where

$$\Delta = (2M + N)^2 + 4r_1^2 + 4r_2^2 - 4(M + N)(r_1 + r_2). \tag{A.84}$$

Showing that $\Delta > 0$ is equivalent to

$$4M^2 + N^2 + 4MN + 4r_1^2 + 4r_2^2 - 4Mr_1 + 4Mr_2 - 4Nr_1 - 4Nr_2 > 0, \tag{A.85}$$

where dividing both sides by 4 leads to

$$M^2 + \frac{N^2}{4} + MN + r_1^2 + r_2^2 - Mr_1 - Mr_2 - Nr_1 - Nr_2 > 0. \tag{A.86}$$

This is equivalent to

$$MN + M(M - N) + \frac{N^2}{4} + MN + r_1^2 + r_2^2 - Mr_1 - Mr_2 - Nr_1 - Nr_2 > 0, \tag{A.87}$$

which can be written as

$$r_2(r_2 - N) + r_1(r_1 - N) + M(N - r_1) + M(N - r_2) + \frac{N^2}{4} + M^2 - MN > 0, \quad \text{(A.88)}$$

leading to

$$(N - r_2)(M - r_2) + (N - r_1)(M - r_1) + \left(M - \frac{N}{2}\right)^2 > 0, \quad \text{(A.89)}$$

which is a sum of positive terms when the inequalities in (5.5) are satisfied. Consequently, equation (5.19) is satisfied when

$$r < \nabla_1 \text{ or } r > \nabla_2, \quad \text{(A.90)}$$

where

$$\begin{aligned} \nabla_1 &= \frac{(2M + N) - \sqrt{\Delta}}{2}, \\ \nabla_2 &= \frac{(2M + N) + \sqrt{\Delta}}{2}, \end{aligned} \quad \text{(A.91)}$$

and $\Delta$ is defined in equation (A.84). However, Lemma 1 bounds the rank r by

$$\max\{r_1, r_2\} \leq r \leq r_1 + r_2. \quad \text{(A.92)}$$

In the following, the idea is to show under which conditions the solution $\nabla_1$ is in the interval $\Big[\max\{r_1, r_2\},\, r_1 + r_2\Big]$. Let us begin by proving the following inequality

$$\nabla_1 < r_1 + r_2. \quad \text{(A.93)}$$

This is equivalent to

$$\frac{2M + N - \sqrt{(2M + N)^2 - 4(M + N)(r_1 + r_2) + 4r_1^2 + 4r_2^2}}{2} < r_1 + r_2, \quad \text{(A.94)}$$

which can be written as

$$(2M + N - 2r_1 - 2r_2)^2 < (2M + N)^2 - 4(M + N)(r_1 + r_2) + 4r_1^2 + 4r_2^2. \quad \text{(A.95)}$$

This leads to the following inequality

$$\begin{aligned} (2M + N)^2 + (2r_1 + 2r_2)^2 - 2(2M + N)(2r_1 + 2r_2) < \\ (2M + N)^2 - 4(M + N)(r_1 + r_2) + 4r_1^2 + 4r_2^2. \end{aligned} \quad \text{(A.96)}$$

Subtracting the term $(2M+N)^2$ from both sides gives

$$(2r_1+2r_2)^2-4(2M+N)(r_1+r_2)<-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2, \quad \text{(A.97)}$$

further expanding the term on the left sides gives

$$\begin{aligned}4r_1^2+4r_2^2+8r_1r_2-4(M+N)(r_1+r_2)-4M(r_1+r_2)<\\ -4(M+N)(r_1+r_2)+4r_1^2+4r_2^2,\end{aligned} \quad \text{(A.98)}$$

which after subtracting the term $-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2$ from both sides leads to

$$8r_1r_2-4M(r_1+r_2)<0. \quad \text{(A.99)}$$

Dividing both sides by 4 gives

$$2r_1r_2-M(r_1+r_2)<0, \quad \text{(A.100)}$$

which can be written as

$$r_1(r_2-M)+r_2(r_1-M)<0. \quad \text{(A.101)}$$

When the inequalities in (5.5) are satisfied the term on the left side is always negative. Therefore the inequality in (A.93) is satisfied. The second part is to show when

$$\nabla_1>\max\{r_1,r_2\}, \quad \text{(A.102)}$$

This is equivalent to

$$\frac{2M+N-\sqrt{(2M+N)^2-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2}}{2}>\max\{r_1,r_2\}, \quad \text{(A.103)}$$

which leads to

$$2M+N-2\max\{r_1,r_2\}>\sqrt{(2M+N)^2-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2}. \quad \text{(A.104)}$$

Raising both sides to the power of two gives

$$(2M+N-2\max\{r_1,r_2\})^2>(2M+N)^2-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2, \quad \text{(A.105)}$$

which is the same as

$$\begin{aligned}(2M+N)^2+4(\max\{r_1,r_2\})^2-4(2M+N)\max\{r_1,r_2\}>\\ (2M+N)^2-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2.\end{aligned} \tag{A.106}$$

Subtracting $(2M+N)^2$ from both sides gives

$$4(\max\{r_1,r_2\})^2-4(2M+N)\max\{r_1,r_2\}>-4(M+N)(r_1+r_2)+4r_1^2+4r_2^2. \tag{A.107}$$

Dividing both sides of the last inequality by 4 leads to

$$(\max\{r_1,r_2\})^2-(2M+N)\max\{r_1,r_2\}>-(M+N)(r_1+r_2)+r_1^2+r_2^2, \tag{A.108}$$

which is equivalent to

$$\begin{aligned}(\max\{r_1,r_2\})^2-(M+N)\max\{r_1,r_2\}-M\max\{r_1,r_2\}>\\ -(M+N)(r_1+r_2)+r_1^2+r_2^2.\end{aligned} \tag{A.109}$$

To simplify the analysis we divide the problem into two cases. Case 1, when

$$\max\{r_1,r_2\}=r_1. \tag{A.110}$$

Making the substitution in (A.109) leads to

$$r_1^2-(M+N)r_1-Mr_1>-(M+N)(r_1+r_2)+r_1^2+r_2^2. \tag{A.111}$$

Subtracting $r_1^2-(M+N)r_1$ from both sides gives

$$-Mr_1>-(M+N)r_2+r_2^2, \tag{A.112}$$

which can be written as

$$Mr_2-Mr_1>r_2^2-Nr_2. \tag{A.113}$$

The last inequality is equivalent to

$$M(r_2-r_1)>r_2(r_2-N). \tag{A.114}$$

Case 2:

$$\max\{r_1,r_2\}=r_2. \tag{A.115}$$

Making the substitution in (A.109) leads to

$$r_2^2 - (M+N)r_2 - Mr_2 > -(M+N)(r_1+r_2) + r_1^2 + r_2^2. \tag{A.116}$$

Subtracting $r_2^2 - (M+N)r_2$ from both sides gives

$$-Mr_2 > -(M+N)r_1 + r_1^2, \tag{A.117}$$

which can be written as

$$Mr_1 - Mr_2 > r_1^2 - Nr_1. \tag{A.118}$$

The last inequality is equivalent to

$$M(r_1 - r_2) > r_1(r_1 - N). \tag{A.119}$$

Combining the results in Case 1 and Case 2 results in (A.102) being satisfied when the following inequality holds

$$M(\min\{r_1, r_2\} - \max\{r_1, r_2\}) > \min\{r_1, r_2\}(\min\{r_1, r_2\} - N). \qquad \square$$

**Theorem 2.** *(Necessary and sufficient condition for beneficial joint recovery) Given two low rank matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M\times N}$ the joint recovery of the two matrices requires less samples that the independent recovery when*

$$M(\min\{r_1, r_2\} - \max\{r_1, r_2\}) > \min\{r_1, r_2\}(\min\{r_1, r_2\} - N), \tag{A.120}$$

*and the rank of the aggregated matrix satisfies*

$$r < \frac{2M + N - \sqrt{(2M+N)^2 - 4(M+N)(r_1+r_2) + 4r_1^2 + 4r_2^2}}{2}. \tag{A.121}$$

*Proof.* The first part of the theorem is included in the proof for Lemma 2. For the second part, i.e., the rank inequality, the same ideas as in the proof for Lemma 2 is followed. Recall that the beneficial joint recovery is defined when

$$\underline{k} < \underline{k}_1 + \underline{k}_2. \tag{A.122}$$

This is equivalent to

$$(2M + N - r)r < (M + N - r_1)r_1 + (M + N - r_2)r_2, \tag{A.123}$$

which is a quadratic inequality. In this context, Lemma 2 shows that when the inequality in (5.30) is satisfied one of the solutions of the quadratic is in the feasible interval for the rank of the combined matrix which is defined by Lemma 1 as

$$\nabla_1 \in \Big[\max\{r_1, r_2\},\, r_1 + r_2\Big], \tag{A.124}$$

where

$$\nabla_1 = \frac{(2M+N) - \sqrt{\Delta}}{2}. \tag{A.125}$$

Therefore, the inequality for the beneficial joint recovery is satisfied when

$$r < \nabla_1, \tag{A.126}$$

which is equivalent to

$$r < \frac{2M + N - \sqrt{(2M+N)^2 - 4(M+N)(r_1+r_2) + 4r_1^2 + 4r_2^2}}{2}. \qquad \square$$

**Theorem 3.** *Let* $\mathbf{M}_1 \in \mathbb{R}^{M\times N}$ *and* $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$*, with* $\operatorname{rank}(\mathbf{M}_1) = r_1$*, and* $\operatorname{rank}(\mathbf{M}_2) = r_2$*. When the locations of the missing entries are sampled uniformly at random with probability* $p_1$ *and* $p_2$ *respectively, the probability of recovery is bounded by*

$$\Pr[E_i] \geq \begin{cases} 1 - e^{-2\frac{(MNp_i - k_i)^2}{MN}}, & \text{for } p_i > \frac{k_i}{MN} \\ 0, & \text{for } p_i \leq \frac{k_i}{MN}, \end{cases} \tag{A.127}$$

*where* $E_i$ *is the event in which the matrix* $\mathbf{M}_i$ *is recovered and* $i \in \{1, 2\}$*.*

*Proof.* When the entries in $\mathbf{M}_1$ and $\mathbf{M}_2$ are uniformly sampled with probability $p_1$ and $p_2$ respectively, the probability of the events $E_1$ and $E_2$ is given by

$$\Pr[E_1] = \sum_{K_j=K_1}^{MN} \frac{(MN)!}{K_j!(MN-K_j)!} p_1^{K_j}(1-p_1)^{MN-K_j}, \tag{A.128}$$

and

$$\Pr[E_2] = \sum_{K_j=K_2}^{MN} \frac{(MN)!}{K_j!(MN-K_j)!} p_2^{K_j}(1-p_2)^{MN-K_j}, \tag{A.129}$$

respectively. Using Hoeffding's inequality for Bernoulli random variables gives

$$\Pr[K_i \leq (p_i - \varepsilon)MN] \leq e^{-2\varepsilon^2 MN}, \tag{A.130}$$

for $\varepsilon > 0$. Therefore, setting

$$\varepsilon = \frac{MNp_i - \underline{k_i}}{MN}, \tag{A.131}$$

for $i \in \{1, 2\}$ results in

$$\Pr[K_i \leq \underline{k_i}] \leq e^{-2\frac{(MNp_i - \underline{k_i})^2}{MN}}. \tag{A.132}$$

Note that in order to satisfy $\varepsilon > 0$ the inequalities hold only for

$$p_i > \frac{\underline{k_i}}{MN}, \tag{A.133}$$

where $i \in \{1, 2\}$. Interestingly, the value $p_i = \frac{\underline{k_i}}{MN}$ is the sampling probability that on average generates a set of $\underline{k_i}$ available observations from the matrix $\mathbf{M}_i$. Also

$$\Pr[K_i \leq \underline{k_i}] + \Pr[K_i > \underline{k_i}] = 1. \tag{A.134}$$

Therefore, the second probability term is

$$\Pr[K_i > \underline{k_i}] = 1 - \Pr[K_i \leq \underline{k_i}], \tag{A.135}$$

which is bounded by (A.132) yielding

$$\Pr[K_i > \underline{k_i}] \geq 1 - e^{-2\frac{(MNp_i - \underline{k_i})^2}{MN}}, \tag{A.136}$$

for $p_i > \frac{\underline{k_i}}{MN}$ and $i \in \{1, 2\}$.

Therefore, the probability of recovering the matrix $\mathbf{M}_i$ when the entries are sampled uniformly at random with probability $p_i$ is bounded by

$$\Pr[E_i] \geq \begin{cases} 1 - e^{-2\frac{(MNp_i - \underline{k_i})^2}{MN}}, & \text{for } p_i > \frac{\underline{k_i}}{MN} \\ 0, & \text{for } p_i \leq \frac{\underline{k_i}}{MN}, \end{cases} \tag{A.137}$$

where $i \in \{1, 2\}$. □

**Theorem 4.** *Let* $\mathbf{M}_1 \in \mathbb{R}^{M \times N}$ *and* $\mathbf{M}_2 \in \mathbb{R}^{M \times N}$, *with* $\operatorname{rank}(\mathbf{M}_1) = r_1$, *and* $\operatorname{rank}(\mathbf{M}_2) = r_2$. *Given the combined matrix* $\mathbf{M} = [\mathbf{M}_1^T \, \mathbf{M}_2^T]^T$ *of rank* $r$, *when the missing entries from* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are sampled uniformly at random with probability* $p_1$ *and* $p_2$, *respectively, the probability of recovering the matrix* $\mathbf{M}$ *is bounded*

*by*

$$\Pr[E_3] \geq \begin{cases} 1 - e^{-\frac{(MN(p_1+p_2)-\underline{k})^2}{MN}}, & \text{for } p_1 + p_2 > \frac{\underline{k}}{MN} \\ 0, & \text{for } p_1 + p_2 \leq \frac{\underline{k}}{MN}, \end{cases} \tag{A.138}$$

*where $E_3$ is the event in which the matrix $\mathbf{M}$ is recovered and is defined in (5.35).*

*Proof.* For the joint recovery case, i.e., the event $E_3$, the number of available entries is $K_1 + K_2$ which is a random variable generated by a Poisson binomial distribution with parameters $2MN$ and $p$ [113], where $p$ is defined in (5.32). In this framework, Lemma 1 in [116] bounds the Poisson binomial cumulative distribution function by the binomial cumulative distribution function with parameters $2MN$ and $\frac{p_1+p_2}{2}$. Therefore the following inequality holds

$$\Pr[K_1 + K_2 \leq \underline{k}] \leq \sum_{K_j=\underline{k}}^{2MN} \frac{(2MN)!}{K_j!(2MN - K_j)!} \left(\frac{p_1 + p_2}{2}\right)^{K_j} \left(1 - \frac{p_1 + p_2}{2}\right)^{2MN-K_j}. \tag{A.139}$$

Using the Hoeffding's inequality for binomial distributions with parameters $2MN$ and $\frac{p_1+p_2}{2}$ yields

$$\Pr[K_1 + K_2 \leq \underline{k}] \leq e^{-\frac{(MN(p_1+p_2)-\underline{k})^2}{MN}}, \tag{A.140}$$

for $p_1 + p_2 > \frac{\underline{k}}{MN}$.

Moreover

$$\Pr[K_1 + K_2 \leq \underline{k}] + \Pr[K_1 + K_2 > \underline{k}] = 1. \tag{A.141}$$

Combining (A.140) and (A.141) leads to

$$\Pr[K_1 + K_2 > \underline{k}] \geq 1 - e^{-\frac{(MN(p_1+p_2)-\underline{k})^2}{MN}}, \tag{A.142}$$

for $p_1 + p_2 > \frac{\underline{k}}{MN}$, which is equivalent to

$$\Pr[E_3] \geq 1 - e^{-\frac{(MN(p_1+p_2)-\underline{k})^2}{MN}}, \tag{A.143}$$

for $p_1 + p_2 > \frac{\underline{k}}{MN}$.

Therefore, the probability of recovering the matrix $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix}$ when the entries in $\mathbf{M}_1$ are sampled uniformly at random with probability $p_1$ and the entries in $\mathbf{M}_2$

are sampled uniformly at random with probability $p_2$ is bounded by

$$\Pr[E_3] \geq \begin{cases} 1 - e^{-\frac{(MN(p_1+p_2)-\underline{k})^2}{MN}}, & \text{for } p_1 + p_2 > \frac{\underline{k}}{MN} \\ 0, & \text{for } p_1 + p_2 \leq \frac{\underline{k}}{MN}. \end{cases} \tag{A.144}$$

□

**Lemma 3.** *Matrix* $\mathbf{M}_i \in \mathbb{R}^{M \times N}$ *with* $\operatorname{rank}(\mathbf{M}_i) = r_i$ *is recovered with probability at least* $q_i$ *if the sampling is uniformly at random with probability* $p_i$ *that is bounded by*

$$p_i \geq \frac{2MN(M+N-r_i)r_i + \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2}, \tag{A.145}$$

*for* $i \in \{1, 2\}$.

*Proof.* The probability of recovering the matrix $\mathbf{M}_i$ is bounded by

$$\Pr[E_i] \geq 1 - e^{-2\frac{(MNp_i-\underline{K_i})^2}{MN}}, \tag{A.146}$$

when $p_i \geq \frac{\underline{K_i}}{MN}$, where $\underline{K_i} = (M+N-r_i)r_i$. In this context, the claim is equivalent to showing that

$$1 - e^{-2\frac{(MNp_i-\underline{K_i})^2}{MN}} \geq q_i, \tag{A.147}$$

which is the same as

$$-2\frac{(MNp_i - \underline{K_i})^2}{MN} \geq \ln(1-q_i). \tag{A.148}$$

Expanding the term on the left side and moving all terms to the left side leads to

$$M^2N^2p_i^2 - 2MN\underline{K_i}p_i + \underline{K_i}^2 + \frac{1}{2}MN\ln(1-q) \geq 0. \tag{A.149}$$

In this context, the value of $\Delta$ for the quadratic is given by

$$\Delta = 4M^2N^2\underline{K_i}^2 - 4M^2N^2\Big(\underline{K_i}^2 + \frac{1}{2}MN\ln(1-q_i)\Big), \tag{A.150}$$

which is further equal to

$$\Delta = -2M^3N^3\ln(1-q_i). \tag{A.151}$$

Note that for $q_i \in (0, 1)$ the log term is bounded by

$$-\infty < \ln(1-q_i) < 0. \tag{A.152}$$

Consequently, $\Delta$ is positive

$$\Delta > 0. \tag{A.153}$$

In this context, the inequality in (A.149) is satisfied when

$$\begin{cases} p_i \leq x_1 \\ p_i \geq x_2, \end{cases} \tag{A.154}$$

where $x_1$ and $x_2$ are the two solutions of the quadratic and are given by

$$x_1 = \frac{2MN\underline{K_i} - \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2}, \tag{A.155}$$

and

$$x_2 = \frac{2MN\underline{K_i} + \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2}. \tag{A.156}$$

However, the probability $p_i$ is bounded by

$$p_i \geq \frac{\underline{K_i}}{MN}. \tag{A.157}$$

The next step in the proof is to show that

$$x_2 \geq \frac{\underline{K_i}}{MN}. \tag{A.158}$$

This yields

$$\frac{2MN\underline{K_i} + \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2} \geq \frac{\underline{K_i}}{MN}, \tag{A.159}$$

which is equivalent to

$$2M^2N^2\underline{K_i} + \sqrt{-2M^4N^4\ln(1-q_i)} \geq 2M^2N^2\underline{K_i}. \tag{A.160}$$

Subtracting the term $2M^2N^2\underline{K_i}$ from both sides leads to

$$\sqrt{-2M^4N^4\ln(1-q_i)} \geq 0, \tag{A.161}$$

which is true for $q_i \in (0,1)$. Consequently, the inequality

$$x_2 \geq \frac{\underline{K_i}}{MN}, \tag{A.162}$$

is true for $q_i \in (0,1)$. Moreover, we can show that the following inequality also holds

$$x_1 < \frac{\underline{K_i}}{MN}. \tag{A.163}$$

This is equivalent to

$$\frac{2MN\underline{K_i} - \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2} < \frac{\underline{K_i}}{MN}, \tag{A.164}$$

which is the same as

$$2M^2N^2\underline{K_i} - \sqrt{-2M^4N^4\ln(1-q_i)} < 2M^2N^2\underline{K_i}. \tag{A.165}$$

Subtracting $2M^2N^2\underline{K_i}$ from both sides leads to

$$\sqrt{-2M^4N^4\ln(1-q_i)} > 0, \tag{A.166}$$

which is true for $q_i \in (0,1)$. Combining (A.154) with (A.157), (A.162) and (A.163) leads to the final solution given by

$$p_i \geq \frac{2MN\underline{K_i} + \sqrt{-2M^3N^3\ln(1-q_i)}}{2M^2N^2}. \qquad \square$$

**Lemma 4.** *Matrix* $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2M\times N}$ *with* $\mathrm{rank}(\mathbf{M}) = r$ *is recovered with probability at least* $q$ *if the matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are sampled uniformly at random with probabilities* $p_1$ *and* $p_2$*, respectively, and the sampling probabilities satisfy*

$$p_1 + p_2 \geq \frac{2MN(M+N-r)r + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \tag{A.167}$$

*Proof.* The probability of recovering the matrix $\mathbf{M}$ is bounded by

$$\Pr[E_3] \geq 1 - e^{-\frac{(MN(p_1+p_2)-\underline{K})^2}{MN}}, \tag{A.168}$$

when $p_1 + p_2 \geq \frac{\underline{K}}{MN}$, and $\underline{K} = (M+N-r)r$. In this context, the claim is equivalent to showing that

$$1 - e^{-\frac{(MN(p_1+p_2)-\underline{K})^2}{MN}} \geq q, \tag{A.169}$$

which is the same as

$$-\frac{(MN(p_1+p_2)-\underline{K})^2}{MN} \geq \ln(1-q). \tag{A.170}$$

Denote by $p_3$ the sum of the two probabilities of sampling i.e., $p_3 = p_1 + p_2$. Then the inequality is equivalent to

$$M^2N^2p_3^2 - 2MN\underline{K}p_3 + \underline{K}^2 + MN\ln(1-q) \geq 0. \tag{A.171}$$

To determine the sign of the quadratic depending on the value of $p_3$, the value of $\Delta$ is analyzed

$$\Delta = 4M^2N^2\underline{K}^2 - 4M^2N^2\underline{K}^2 - 4M^3N^3\ln(1-q), \tag{A.172}$$

which is equivalent to

$$\Delta = -4M^3N^3\ln(1-q). \tag{A.173}$$

Consequently $\Delta > 0$ for $q \in (0,1)$. Therefore, the inequality in (A.171) is satisfied when

$$\begin{cases} p_3 \leq y_1 \\ p_3 \geq y_2, \end{cases} \tag{A.174}$$

where

$$y_1 = \frac{2MN\underline{K} - \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}, \tag{A.175}$$

and

$$y_2 = \frac{2MN\underline{K} + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \tag{A.176}$$

However, the probability $p_3$ is bounded by

$$p_3 \geq \frac{\underline{K}}{MN}. \tag{A.177}$$

In this context, the next step is to show that

$$y_2 \geq \frac{\underline{K}}{MN}. \tag{A.178}$$

This is equivalent to

$$\frac{2MN\underline{K} + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2} \geq \frac{\underline{K}}{MN}, \tag{A.179}$$

which can be written as

$$2M^2N^2\underline{K} + \sqrt{-4M^4N^4\ln(1-q)} \geq 2M^2N^2\underline{K}. \tag{A.180}$$

Subtracting $2M^2N^2\underline{K}$ from both sides leads to

$$\sqrt{-4M^4N^4\ln(1-q)} \geq 0, \tag{A.181}$$

which is true for $q \in (0,1)$. Moreover, the following inequality also holds

$$y_1 < \frac{\underline{K}}{MN}. \tag{A.182}$$

This is equivalent to

$$\frac{2MN\underline{K} - \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2} < \frac{\underline{K}}{MN}, \tag{A.183}$$

which is the same as

$$2M^2N^2\underline{K} - \sqrt{-4M^4N^4\ln(1-q)} < 2M^2N^2\underline{K}. \tag{A.184}$$

Subtracting $2M^2N^2\underline{K}$ from both sides leads to

$$\sqrt{-4M^4N^4\ln(1-q)} > 0, \tag{A.185}$$

which is true for $q \in (0,1)$ and therefore the inequality in (A.182) holds. Combining the solutions in (A.174) with the condition in (A.177) and the results in (A.178) and (A.182) leads to the final solution given by

$$p_3 \geq \frac{2MN\underline{K} + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \tag{A.186}$$

Coming back to the original notation gives

$$p_1 + p_2 \geq \frac{2MN\underline{K} + \sqrt{-4M^3N^3\ln(1-q)}}{2M^2N^2}. \qquad \square$$

**Lemma 5.** *Let* $\psi \in [0,1]$ *and* $\mathbf{A} \in S_+^M$*, then the matrix* $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{A} & \psi\mathbf{A} \\ \psi\mathbf{A} & \mathbf{A} \end{bmatrix} \in S_+^{2M}$.

*Proof.* Using the Schur complement [117] the matrix $\mathbf{\Sigma}$ is positive semi-definite if and only if the matrix $\mathbf{A} - \psi\mathbf{A}\mathbf{A}^{-1}\psi\mathbf{A}$ is positive semi-definite. Consequently, the matrix $\mathbf{\Sigma}$ is positive semi-definite if and only if $(1-\psi^2)\mathbf{A}$ is positive semi-definite. However, the matrix $\mathbf{A}$ is positive semi-definite and therefore, when $\psi \in [0,1]$, the matrix $(1-\psi^2)\mathbf{A}$ is positive semi-definite. □